DOGUS UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES

APPLICATION OF TEXT MINING ON IT INCIDENT MANAGEMENT SYSTEMS

GRADUATE THESIS

ERDAL SEVER İŞCEN

200891001

ADVISOR
ASSIST. PROF. DR. M. ZAHİD GÜRBÜZ

Istanbul, 2019

DOGUS UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF COMPUTER AND
INFORMATION SCIENCES

APPLICATION OF TEXT MINING ON IT INCIDENT MANAGEMENT SYSTEMS

GRADUATE THESIS

ERDAL SEVER İŞCEN

200891001

ADVISOR
ASSIST. PROF. DR. M. ZAHİD GÜRBÜZ

Istanbul, 2019

	YÜKSEK LİSANS TEZ SINAV TUTANAĞI	*Doküman No*	FR.1.26
		Yürürlük Tarihi	1.11.2017
		Revizyon Tarihi	1.11.2017
		Revizyon No	1
		Sayfa	1 / 1

SOSYAL BİLİMLER / FEN BİLİMLERİ ENSTİTÜSÜ

Tarih : 13/09/2019

Anabilim/Anasanat Dalı : Department of Computer and Information Sciences

Öğrencinin Adı Soyadı :Erdal Sever İŞCEN

Öğrenci No :200891001

Tez Danışmanının Adı Soyadı :Dr.Öğretim Üyesi M. Zahid GÜRBÜZ

İkinci Tez Danışmanının Adı Soyadı :-

Tezin Başlığı :Application of Text Mining on IT Incident Management Systems

Doğuş Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği'nin 32.Maddesi uyarınca yapılan değerlendirmeler sonunda:

☑ **tezin kabul edilmesine** □ **tezde düzeltme verilmesine** □ **tezin reddedilmesine**

[✓ oy birliği] / oy çokluğu ile karar verilmiştir.Gereği için arz olunur.

Danışman Üye

Dr.Öğr.Üyesi M.Zahid Gürbüz

Üye **Üye** **Üye**

Dr.Öğrt.Üys.Aysun GÜRAN

Üye

Prof.Dr.İbrahim Emiroğlu

Anabilim/Anasanat Dalı Başkanı Onayı:

Dr.Öğ.Üye Yasemin Karagül

YEMİN METNİ

Yüksek Lisans tezi olarak sunduğum “***Application of Text Mining on IT Incident Management Systems***” adlı çalışmanın, tarafımdan, akademik kurallara ve etik değerlere uygun olarak yazıldığını ve yararlandığım eserlerin kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve bunu onurumla doğrularım.

13/09/2019

Erdal Sever İŞCEN

ACKNOWLEDGEMENTS

In the preparation of this project, my efforts alone would not be enough to succeed without the support I gained.

I would like to thank my advisor M. Zahid Gürbüz for all his patience guidance throughout all the phases of this study. With his advisory skills and knowledge on the study, we were able to define clear goals and roadmaps for this project. Thanks to the hours he stayed at the university after his class hours waiting for me to follow-up on the progress, we have always been on track.

I also have to thank my manager in Medtronic, Massimiliano Pepe, for his strong support on this project. He has always kept track of the status of my project and encouraged me to focus on it. By setting this project as one of my development goals he always helped to keep my motivation up.

Last but not least, I owe my parents for their endless faith in me. They always wanted for my success on this project without putting any pressure on me.

Istanbul, 2019 Erdal Sever İşcen

ABSTRACT

With the increase of data stored in computer systems worldwide, gaining knowledge from data became more computer dependent. Data alone is not very valuable unless there's knowledge extracted from data. This is the reason that machine learning has become a quite important topic recently. Text classification is a machine learning task which aims on classifying documents based on their content and that is the method used in this study.

This study focuses on the task of classifying documents on the IT service management, mainly incident management, tools. IT service management and incident management is a hot topic in every company which serves IT services and they require human effort to manage. In ITIL framework for IT service management, it's always useful to link the incidents with configuration items, in other words the assets or components necessary to deliver IT services, and this task is managed manually by IT support technicians in many IT service management tools. The aim of the study is to remove this manual linking step by applying text classification methods and instead to provide an automatic assignment of CI's to incidents.

Four text classification methods, Naïve Bayes Multinomial, k-Nearest Neighbor, Support Vector Machine and C4.5 decision tree classifiers, are used on three different sets of incidents extracted from the same database by using different filters. The impact of some pre-processing steps is compared on different sizes of datasets.

It's possible to conclude from the experiments that this study achieved an acceptable accuracy on all sizes of datasets though some methods, like SVM on Dataset B, may only be used as nice-to-have as the results are not perfect. Using a model with 64.41% accuracy as a primary solution might cause time and value loss.

Another conclusion from the experimental results is that SVM performs well even in such complex datasets while the classifiers like Naïve Bayes Multinomial and kNN are very sensitive to noise.

Learning speed of each classifier in the evaluations has been monitored and proved that C4.5 could be problematic from speed perspective on complex datasets. Naïve Bayes Multinomial and kNN, however, does not require any time to learn because of their distance and probability approaches

Keywords: Text Classification, Incident Classification, Ticket Classification, Ticket Categorization, Incident Categorization, Text Categorization,

ÖZET

Bilgisayar sistemlerinde saklanan veri hacmi gün geçtikçe hızlı bir şekilde artmakta ve veri boyutu arttıkça veriyi yorumlaması için bilgisayar sistemlerine duyulan ihtiyaç da aynı şekilde artmaktadır. Verinin bir anlam içermesi için yorumlanıp bir bilgiye ulaşılması gerekiyor. Günümüzde bilgiye ulaşmak için makine öğrenmesi teknikleri üzerinde oldukça yoğun çalışmalar yapılmaktadır. Bir makine öğrenmesi yöntemi olan metin sınıflandırma da metin içeren dökümanların içeriğinden bir bilgiye veya sonuca çıkmayı hedeflemektedir.

Bu çalışmada metin sınıflandırma yöntemlerinin BT hizmet yönetimi, özellikle de olay yönetimi, sistemlerindeki metin içerikli kayıtları sınıflandırmak için kullanımı değerlendirilmiştir. Günümüzde, şirket içi veya şirket dışı olsun, BT hizmeti sağlayan şirketlerde BT hizmet yönetimi son derece önem arz eden bir konu haline gelmiştir ve bunun için dünya çapında geçerli belirli standartlar takip edilmektedir. ITIL, dünya çapında geçerli olan BT hizmet yönetimlerinden biridir ve ITIL çerçevesine göre yapılandırma öğeleri, yani BT hizmeti vermek için gerekli olan bileşenlerin tamamı, ve olaylar birbiriyle ilişkilendirilmelidir. Genel olarak kullanılmakta olan BT hizmet yönetimi sistemlerinde bu ilişkilendirme işlemi BT teknisyenlerinin eforlarıyla, elle yapılmaktadır. Yapılan deneylerde elle yapılan bu ilişkilendirme işleminin metin sınıflandırma metotları yardımıyla otomatik bir hale getirilmesi hedeflenmiştir.

Çalışma kapsamındaki uygulamalarda aynı BT hizmet yönetimi sisteminden elde edilen üç farklı veri kümesi kullanılmıştır. Buradaki amaç aynı sınıflandırma metotlarının farklı veri kümelerinde nasıl performans gösterdiklerini gözlemlemektir. Sınıflandırma metodu olarak Naïve Bayes Multinomial, k-En Yakın Komşu, Destekçi Vektör Makineleri ve C4.5 karar ağacı sınıflandırıcıları kullanılmıştır ve bu metotlar farklı önişleme aşamalarıyla birlikte çalıştırılarak önişleme aşamasının sonuca etkisi değerlendirilmiştir.

Sınıflandırma işlemleri sonucunda her veri kümesinden kabul edilebilir bir performans elde edilmiştir fakat örnek olarak B veri kümesinden maksimum 64.41%

doğruluk elde edilmiştir ve bu derece gerekli bir implementasyon olarak değil; ancak bir öneri sunma fonksiyonu gibi ek bir fayda sağlayacak fonksiyon olarak değerlendirilebilir.

Uygulama sonuçlarından çıkan bir başka yorum ise Destekçi Vektör Makineleri sınıflandırıcısının B veri kümesi gibi daha komplike veri kümelerinde bile kabul edilebilecek bir sonuç sağladığıdır. Öte yandan Naïve Bayes Multinomial ve k-En Yakın Komşu algoritmalarının veri kümesindeki değişikliklere veya gürültüye daha hassas oldukları gözlemlenmiştir.

Öğrenme süreleri açısından ise yapılan değerlendirmede ise C4.5 karar ağacı metodunun öğrenme aşamasının son derece uzun sürebildiği ve Naïve Bayes Multinomial, k-En Yakın Komşu algoritmalarının istatistiksel ve mesafe ölçümleri sayesinde öğrenme aşamasında zaman harcamadığı gözlemlenmiştir.

Anahtar Kelimeler: Metin Sınıflandırması, Döküman Sınıflandırması, Metin Madenciliği, Olay Sınıflandırması, Talep Sınıflandırması

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACRONYMS

IT	: Information Technology
ITSM	: Information Technology Service Management
ITIL	: Information Technology Infrastructure Library
COBIT	: Control Objectives for Information and Related Technology
ISO	: International Organization for Standardization
IEC	: International Electrotechnical Commission
CI	: Configuration Item
kNN	: K-Nearest Neighbor
SVM	: Support Vector Machine
SLP	: Service Level Package
SDP	: Service Design Package
BCP	: Business Continuity Plan
CMDB	: Configuration Management Database
CSI	: Continual Service Improvement
TP	: True Positive
FP	: False Positive
NBM	: Naïve Bayes Multinomial

1. INTRODUCTION

Today, especially with the improvements on the storage/memory sizes, processing speeds and network bandwidths, dependency on computer systems has been increased daily operations. Humans can be replaced without much concern but it’s a very critical case to replace a computer system and the data stored in such a system. This is why, Information Technology (IT) has become a key subject in every industry and this led the industry to build worldwide standards in IT Service Management (ITSM). ITIL, COBIT, ISO/IEC are some of the well-known IT Service Management frameworks.

This study focuses on ITIL framework which has the concept of Configuration Item in its scope. “Configuration Item (CI) is defined as any asset, service component, or other item that is, or will be, under the control of Configuration Management. Examples of CIs include hardware and software, applications, network infrastructure, contracts and even people” (Orand & Villarreal, 2011, p.185).

As stated above, with the improvements on the capability of information technology, the data stored in the systems has radically increased and this is the main reason for a need of more effort on discovering knowledge from data. In simplest terms, more data to read means more time required to make use of it. Because of that, the study of machine learning has become quite important recently. Data mining, as a machine learning technique, is the process of discovering knowledge from large amount of data. Text mining is an application of data mining methods on text data and there are different types of text mining applications such as text categorization, clustering and information extraction.

The aim of this study is to classify the IT incidents, which have been created by the end-users, based on their CI’s. In most of the ITSM applications, this requires a manual effort from a support technician to identify the CI. By definition, an incident is an unplanned interruption to an IT service or reduction in the quality of an IT service. With the application of text mining techniques, specifically Text Categorization, it is expected to eliminate the need for the manual effort for classifying the IT incident tickets, or support requests based on their CI’s.

The evaluations of this study have been made based on real data, taken from the incident management tool ServiceNow that is used in one of the largest medical device companies.

The next sections will respectively explain about similar studies that have been published in section 2, the concepts of IT service management in section 3, text mining as a machine learning study in section 4 and the details of our experiments in section 5. In section 6 as a conclusion, a solution is expected based on the accuracy results to automate the task of configuration item selection and eliminate the human effort required for the selection.

2. RELATED WORK

In the study of Zinner et.al (2015), feedback for IT incidents are classified using the similar method but the content to classify instances is different. In their experiements, terms should be linked to two classes and the relationship will depend on the negativity or positivity while our study aims to link the terms with the technical subjects, Configuration Items in ITIL terminology, which are created by the IT teams. The results of their experiments are based on different parameters such as precision, recall and F1 score due to the imbalanced dataset and low amount of classes. However, the class distribution in our dataset, even though there are many classes, can be considered more uniform compared to the data set Zinner et.al (2015) used and unlike their measurements, this study focuses on accuracy as the result parameter. They had the conclusion in their study that Support Vector Machine algorithm outperformed K-Nearest Neighbor and Naïve Bayes algorithms.

Another similar study in this field was managed by Altintas & Tantug (2014). Their research aimed on categorizing over 10000 IT incidents based on their categories with the total of 4 categories and subcategories with the total of 20 subcategories based on the data taken from the Issue Tracking System of Istanbul Technical University. Different from this study, they worked on Turkish dataset. The results proved that accuracy of algorithms depend on the training datasets. SVM proved to be the best performer on larger training datasets while Naïve Bayes and Decision Tree methods performed better in smaller training datasets. Another point was that the study achieved quite high, over 90% accuracy, on classifying categories compared to classifying subcategories because the number of classes in categories is less than the number of subcategories. This also leads us to a point that it’s possible to achieve higher accuracy when working on more distinctive classes.

A different approach on the same problem was applied by Beneker & Gips (2017). In contrast to the other studies stated above and ours, they used an unsupervised approach to the problem and used clustering algorithms such as k-means and Non-negative Matrix Factorization for clustering the tickets without any labeled training sets. The outcome from k-means and Non-negative Matrix Factorization showed between 60% and 70% similarity which varies based on the number of clusters to extract. The less the number of clusters they set, the higher similarity the results had.

Sakolnakorn, Meesad & Clayton (2008) worked on classifying IT incidents based on the resolver group by using decision tree algorithms and achieved over 90% accuracy with all decision tree algorithms on assigning the incidents to the resolving groups. They worked on a dataset of 14440 documents with 5 classes. The study is a proof of how effective decision trees are when working with a dataset that is not complex.

Agarwal, Sindhgatta & Sengupta (2012) worked on classifying IT tickets based on the resolver group by proposing a new approach, SmartDispatch, in order to overcome SVM’s flaws. Instead an approach with a combination of SVM and Discriminative Term Approach is proposed.

Agarwal et.al (2017) made an approach for classifying IT incidents from different perspectives such as symptom diagnosis, root cause diagnosis and action extraction. They applied their own approach and SmartDispatch, which was already proposed by Agarwal & Sindhgatta & Sengupta (2012), in the same dataset and their results were at least as good as SmartDispatch.

There are many other problems where text classification focuses for a solution. Pratama & Sarno (2015) used classification algorithms kNN, Naïve Bayes and SVM for personality classification based on Twitter posts with the use of multi-label approach. Naïve Bayes outperformed other algorithms in their study.

3. IT SERVICE MANAGEMENT

Today, IT plays an important role in most of the organizations because the operations heavily depend on information systems. As a simple real life example, consider a supermarket: Every transaction, such as orders, sales, demand, logistics, of a product has to be tracked and logged. Of course it’s technically possible to manage all these without IT systems but it would require lots of effort, resources and time. With the help of information systems it’s much easier to process, compute, store data; access data between remote locations and generate reports and it’s a matter of seconds to perform each of these tasks.

There has always been a need to standardize the way of managing IT systems just like any other operation. For this need, the topic of IT Service Management has become popular for publishing good practices, in other words standards derived from best practices, in this field. Without a formal discipline, it would be difficult to adapt the same IT services to different organizations. Also, it’s a necessity to measure the quality of IT services in similar terms, just like speaking the same language between parties. Some of the well-known IT Service Management frameworks are ITIL, COBIT, ISO/IEC. While all these frameworks focus on standardization of IT services, there are differences in their approaches. Gehrmann (2012) published a survey in his study showing that ITIL, COBIT and ISO are the top frameworks adopted globally and there had been an increase on the interest on these frameworks compared to past. Another survey published by Forbes in Comptia (2017) shows that 47% of the senior executives, who contributed to the survey, adopt ITIL and 36% adopt COBIT in their organizations.

Because ITIL is the framework that is used as the guidelines for the IT operations in the company, which the dataset belongs to, this study focused on the application ITIL framework.

3.1 ITIL

Information Technology Infrastructure Library (ITIL) is a framework of IT service management that is commonly known and adopted in most of the companies worldwide. Developed in 1980s by the Central Computer and Telecommunications in the UK in order to improve how to manage IT, ITIL has been updated with different modifications over time. The most recent version of ITIL framework is ITIL v4, published in February 2019. However, ITIL v4 is quite new and it requires time to transition processes from ITIL v3 approach. This is why we will focus on ITIL v3, which has been published in 2007; updated in 2011. It's a goal for most of the organizations to switch to ITIL v4 framework but we can consider ITIL v3 as the most common framework for now. ITIL related content in this study is based on ITIL v3 standards. Gil-Gomez, Oltra-Badenes & Adarme-Jaimes (2014) stated in their study that ITIL is a framework that can successfully be applied in many other service providing operations in order to manage the way of providing services.

ITIL v3 consists of 5 core volumes and can be visualized as in Figure 1:

- Service Strategy
- Service Design
- Service Transition
- Service Operation
- Continual Service Improvement

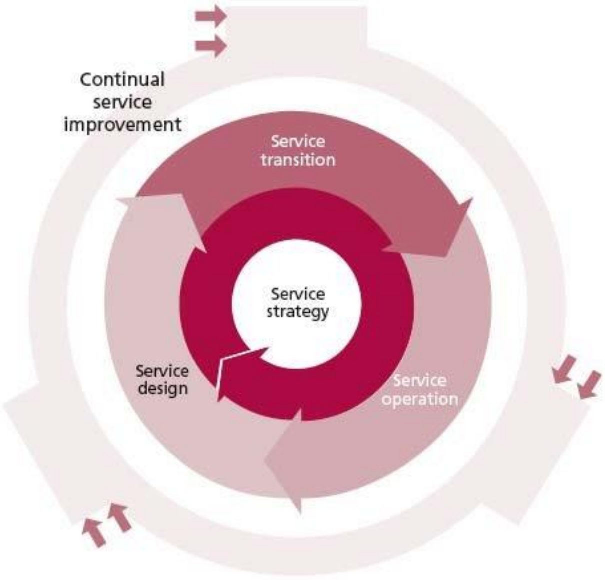


Figure 3.1 ITIL Service Lifecycle
Source: Cartlidge, A., Rudd, C., Smith, M., Wigzel, P., Rance, S., Shaw, S., & Wright, T. (2012). An Introductory Overview of ITIL® 2011. London: The Stationery Office

Service Transition and Service Operation phases will be the main subjects of this study as Incident Management and Configuration Management are in scope of Service Transition and Service Operation respectively.

3.1.1 Service Strategy

In most operations, a strategy must be defined for the long term goals. It is no different in IT. Service Strategy is the core component in ITIL service lifecycle and is responsible for management level decisions such as defining services and the value they create within financial targets, target audiences for services. Technical knowledge is not a main component in Service Strategy; it's the value what's more important in Service Strategy. Service Level Package (SLP), documentation of business and process requirements, is the main output of Service Strategy. As stated in Orand & Villarreal (2011), some of the processes in the Service Strategy can be listed as below:

- Service Portfolio Management: The process of representing and documenting all services that are developed, transitioned, operational or retired.
- Demand Management: Collection of activities for detecting and influencing customer demands based on the services that are provided.
- Financial Management: The process of managing financial inputs and outputs in order to fund the IT services provided by the organization.

3.1.2 Service Design

Service Design takes the SLP as an input from Service Strategy and aims to provide a Service Design Package (SDP), which defines the aspects of an IT service, to the Service Transition phase. The basic goal of this process is to design the services such as management and system tools, service management metrics and processes. This is the phase that moving the strategy and designing activities to the operational tasks take place. Below are the sub-processes of Service Design:

- Service Level Management: Responsible for measuring the services and their performances.
- Service Catalog Management: Similar to service portfolio, a service catalog must be maintained in ITIL adopted operation. A portfolio contains the set of services and represents the commitments, investments on these services. Service catalog, on the other hand, is the set of services that can be provided and supported by the organization. (Arcilla, Calvo-Manzano & San Feliu, 2013)
- Availability Management: Focuses on the availability of the services. Availability is a measurable value which is also known as uptime.
- Capacity Management: Main objective of capacity management is to make sure that the capacity of services meet the current and future needs of the customers, or users.
- Information Security Management: Is the measure of securing information effectively through availability, confidentiality, integrity and authenticity.

- Supplier Management: Management of vendors or suppliers in order to make sure that the organization obtains the services in the agreed level.
- IT Service Continuity Management: Mainly focuses on the BCP’s and disaster recovery plans for the business and operation continuity.

3.1.3 Service Transition

Service Transition phase depends on the SDP’s developed in the Service Design phase and aims to prepare the service for operation. Service Transition Package, which contains metrics, service levels and procedures, is the output of this phase.

Some of the processes of Service Transition phase can be listed as:

- Change Management: Process of managing Change Requests. A Change Request is necessary in ITIL framework in order to make changes in the services or Configuration Items. Ignoring this process might lead to inconsistencies between the operations and documentations.
- Configuration Management: This is the process which helps to control the infrastructure. IT assets, could be hardware or software, of an organization must be stored and documented properly as Configuration Items (CI). Without such a documentation, an IT organization can be considered empty, without any asset. Another reason to document CI’s well is that other processes such as Change Management and Incident Management depend on CI’s. In case of a Change Request, it’s necessary to list the CI’s that will be affected so that the organization would be aware of the link between a Change Request and the CI. If the change request fails, it will then be visible what CI’s will be impacted.
- Release Management: Focuses on releasing a service into production with proper preparations and planning such as training, documentation, testing, roll back plans.
- Knowledge Management: Discipline of making sure that the knowledge in an organization is maintained, documented and updated properly.

3.1.4 Service Operation

The customer receives the value from an IT service in Service Operation phase. Service Desk, Technical Management, IT Operations Management, Application Management are some of the functions operating in Service Operation. Sub processes of this phase are:

- Event Management: The documentation of change of states of CI's in an organization. Can be considered as notifications.
- Incident Management: An incident is an unplanned interruption or loss of quality on a service or CI. Our evaluation mainly focuses on Incident Management because we work on the incidents that have been reported. An Incident must be documented in a proper way in order to link them other components such as CI's, other incidents, services or vendors.
- Request Fulfillment: The process of managing requests for standard changes such as adding software to computers, requesting a toner or any hardware that has none, or very small, impact on a service.
- Problem Management: Focuses on the root cause analysis of incidents. The aim of this process is to reduce the amount of incidents by eliminating the root causes.
- Access Management: Responsible for managing access requests from users and keeping track of the existing access in case there are modifications due to a change of user's role in the organization.

3.1.5 Continual Service Improvement

All the stages we mentioned above should be covered and improved in a continuous and iterative way. It is possible that the objectives of an organization change and for that there's always a need to review the existing processes by asking the questions in Figure 3.2:

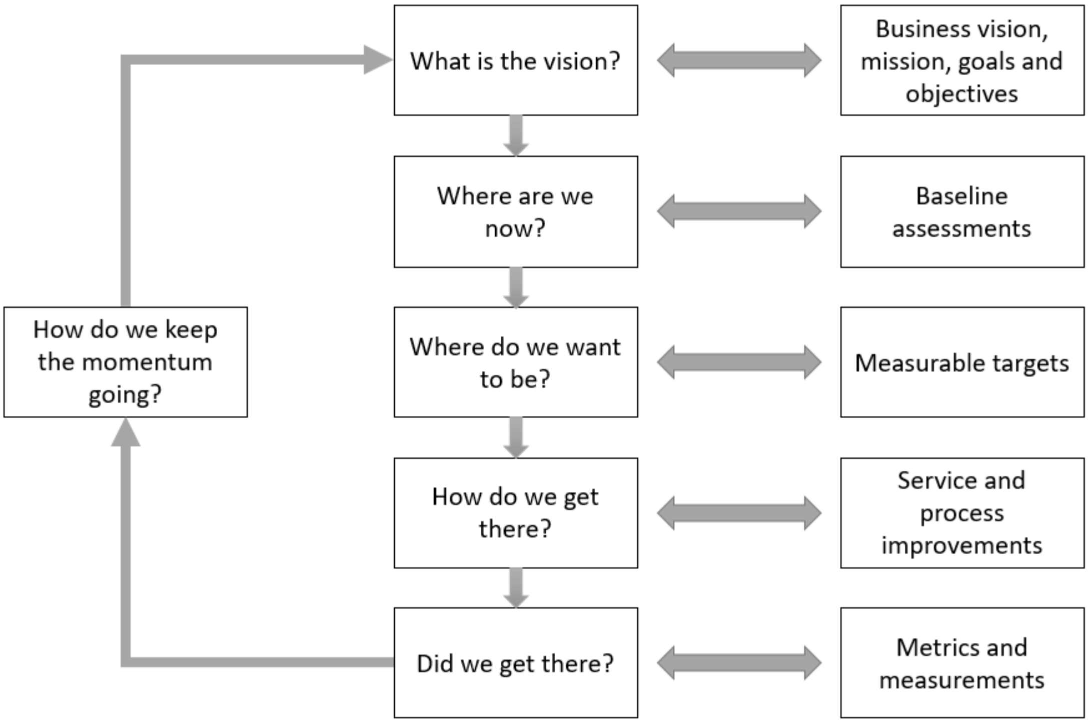
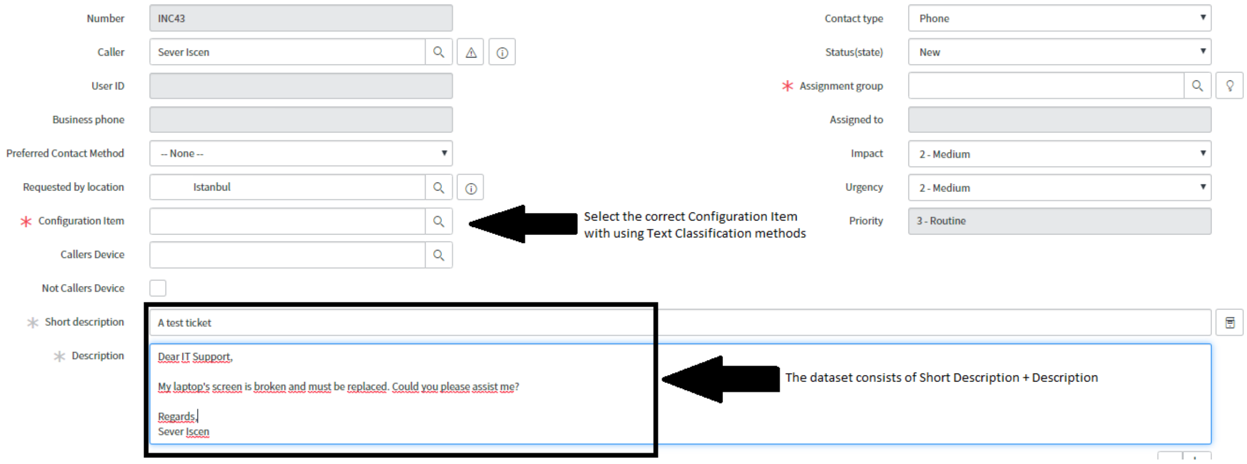


Figure 3.2 Questions to ask in CSI

3.2 ServiceNow as an IT Service Management Tool

As stated in the previous section, there are many processes to consider in IT Service Management. In order to manage services using these processes, an ITSM software is always useful. ServiceNow is one of the ITSM applications known worldwide and the data used for this study was extracted from ServiceNow Incident Management process. ServiceNow covers many more processes too however because the objective of this study is based on the incidents that are reported, Incident Management process will be the main process to focus in this study. In Figure 3.3, a sample incident is shown.



Number	INC43	Contact type	Phone
Caller	Sever Iscen	Status(state)	New
User ID		* Assignment group	
Business phone		Assigned to	
Preferred Contact Method	-- None --	Impact	2 - Medium
Requested by location	Istanbul	Urgency	2 - Medium
* Configuration Item		Priority	3 - Routine
Callers Device			
Not Callers Device	□		
* Short description	A test ticket		
* Description	Dear IT Support, My laptop's screen is broken and must be replaced. Could you please assist me? Regards, Sever Iscen		

Select the correct Configuration Item with using Text Classification methods

The dataset consists of Short Description + Description

Figure 3.3 Sample ticket in ServiceNow

4. DATA MINING

With the age of digitalization, the amount of data stored in computer systems has increased dramatically all around the world and it requires more effort to gain knowledge from the data as the size of data grows. It is now preferred that computer systems gain this knowledge from data and summarize the outcomes from it for human use. Machine Learning is the concept of learning by computer systems based on the data it processes. Data mining, as part of machine learning, aims to extract knowledge from data by using statistical methods.

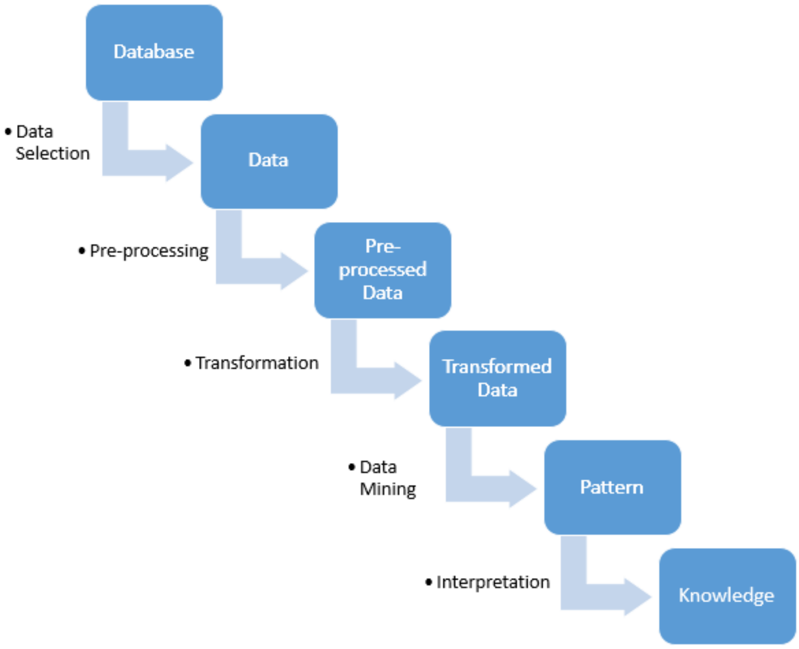


Figure 4.1 From data to knowledge. *Source: Şeker, Ş (2013). İş zekası ve veri madenciliği. Istanbul*

In machine learning, there are four types of learning methods as listed below.

4.1 Unsupervised Learning

Unsupervised learning is the way of learning where the data available has no label. In this type of learning, it is up to the system to create its own labels and extract information by using those labels. Clustering algorithms such as K-Means and Non-Negative Matrix Factorization are some of the unsupervised algorithms.

4.2 Semi-supervised Learning

Using a combination of both labeled and unlabeled data in a problem is in the scope of semi-supervised learning. In this approach, labeled data is used to learn the model of the existing classes and the unlabeled data is used to discover the boundaries between the classes. Some examples of semi-supervised learning can be genetic sequencing, web page classification where we don’t have the whole data labeled.

4.3 Active Learning

Active learning is the method of learning which requires user’s involvement to learn the model. The main aspect of this method is to improve the quality by relying on experts’ knowledge.

4.4 Supervised Learning

The main focus of this study, supervised learning, is the learning method where the training data is already labeled. Labeled training data has been available in this study. Regression and classification are the main areas of uses in supervised learning. The classifications algorithms, which are part of supervised learning, Naïve Bayes Multinomial, SVM, C4.5 and kNN are used in this study. The way these four classifiers work has been explained in the next sections.

4.4.1 Naïve Bayes Multinomial

Naïve Bayes is considered as an effective statistical classifier based on the accuracy and speed, especially in large databases. Naïve Bayes is one of the fastest classifiers on training the system out of the other three algorithms selected for our experiments and the results in the evaluation section prove this. The main aspect of Naïve Bayes classifier is class conditional independence, meaning that the terms are independent from each other and has no impact on the result based on their co-occurrence. A probability of an instance X of belonging to the class C_i can be formulated as below.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (4.1)$$

Where P(X|C_i) stands for the probability of class C_i containing the terms of X, P(C_i) is the probability of class C_i in all the training set and P(X) is the probability of terms X in all the training set regardless of class.

Naïve Bayes Multinomial (NBM) is another method of classification algorithm which is optimized for text classification tasks. NBM uses the word frequency, or term weight, in order to decide on the probability of a document. Just like Naïve Bayes, NBM is based on term independence: Position of a term and length of a document have no impact on the classification task (McCallum & Nigam, 1998).

4.4.2 Support Vector Machine

SVM classifier aims to create its own boundaries, or hyperplanes, when learning a model. They work well on both linear and non-linear data. SVM works on finding the hypothesis with the lowest true error. Because the classifier works on building its own hyperplanes for distinguishing classes even in the most complex datasets, the training time is quite higher than many other classifiers while the accuracy results are usually competitive against other classifers. SVM can also be used in predictions.

In the process of identifying the hyperplane for separating classes, it’s possible to generate more than one hyperplane. However, SVM aims to find the maximum marginal hyperplane which is unique and the most distinguishing boundary.

An example dataset with two attributes A1 and A2 is displayed in Figure 3.2. In the graph on the left side, a hyperplane is suggested but apparently it does not the largest possible margin to the closest instances from different classes. However on the right side of Figure 4.2, a hyperplane with a larger margin is suggested and it definitely is more distinctive. The goal of SVM is to find out this imaginary hyperplane with the largest margin possible to build another space for distinguishing documents.

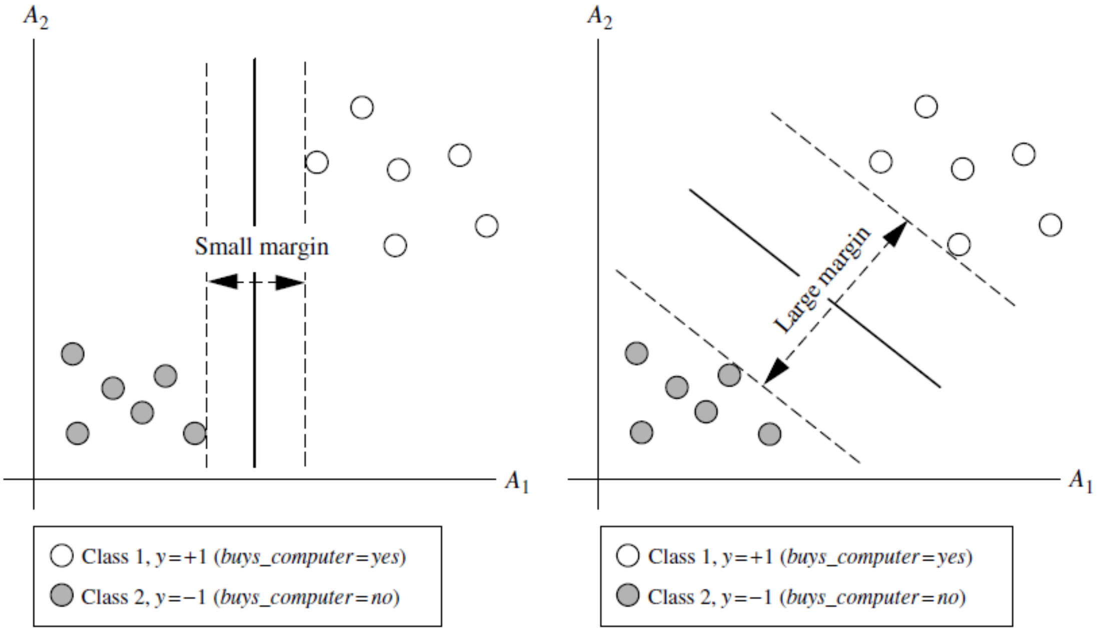

A_2
Small margin
A_1
○ Class 1, $y=+1$ (*buys_computer=yes*)
● Class 2, $y=-1$ (*buys_computer=no*)
A_2
Large margin
A_1
○ Class 1, $y=+1$ (*buys_computer=yes*)
● Class 2, $y=-1$ (*buys_computer=no*)

Figure 4.2 Different margins on the same dataset. Source: Han, J., Kamber, M., Pei, J. (2011) Data Mining: Concepts and Techniques (3rd edition). San Francisco: Morgan Kaufmann.

Joachims (1998) proved in his experiments that SVM achieves good accuracy on text classification without the need of feature selection thanks to its efficient generalization capability even in high dimensional feature spaces. Also, SVM is proved to work well on multi-label classification tasks, Qin & Wang (2009) were able to achieve high performance on multi-label classification by proposing approaches based on SVM.

4.4.3 C4.5 Decision Tree

In a classification task, decision trees build a tree, with nodes, branches and leaves, in order to find out the class based on the terms of the instance. Each node represents the decision based on a specific term. Branches are the connections between nodes which are leading to another node. In other words, branch is the answer of a single test at a node. The leaves of a tree represent the class decision which the end points of a decision tree.

In order to build the tree, decision trees use different attribute selection methods. C4.5 specifically uses Gain Ratio as the attribute selection method.

4.4.4 K-Nearest Neighbor

kNN is a distance based classifier which is based on some similarity measures such as Cosine or Euclidian. The kNN evaluations in this study adopted the Euclidian distance metric. The Euclidean distance between two points A and B can be formulated as below in (4.2):

$$distance(A, B) = \sqrt{\frac{\sum_{i=1}^{m}(x_i - y_i)^2}{m}} \tag{4.2}$$

kNN is a simple classifier to run in experiments but it has its own difficulties. Although there’s no training time required for kNN, classification task takes much longer compared to other classifiers; results in Table 5.5 already prove this. Another challenge with kNN is to find the optimal k value. The class decision in the classification task is based on the closest k values to the document. Although it’s commonly suggested to use k as the square root of the dataset size k=7 is used in the experiments as it gave the highest accuracy compared to the other iterations tried in the efforts of finding the optimal k value.

5. EVALUATION

Automatic categorizing CI’s of the incidents taken from the ITSM software, ServiceNow is the goal of the evaluations. With the reporting capabilities of ServiceNow, exportation of incidents was an easy task with specific filters such as language, region and the source to create our dataset. After parsing the export and converting the csv report to an arff file format, dataset was loaded to Weka.

Weka, Waikato Environmant for Knowledge Analysis, is a data mining application developed at the University of Waikato. It was originally developed for processing agricultural data. Predicting the internal bruising of apples, quality of mushrooms are some agricultural examples of using Weka in late 90s (Frank et.al, 2009, p.1275). With the increased focus on machine learning study, Weka has become a tool for many other applications such as Neuro-lingiustic programming, bioinformatics and many other industries. As part of the data mining methods, text mining is also a field which we can use Weka for.

The experiments in this study were performed on a workstation with Intel Core i7-3770K CPU @ 3.5GHz and 16GB of memory on a Windows 10 (64-bit) installation.

Accuracy, which is the number of correctly classified instances in the test split of the dataset, has been selected as the primary metric to measure the performance of the classifiers.

In the next sections, we will explain how the experiments were operated using Weka and list the outcomes of the experiments.

5.1 Dataset

The data is exported from ServiceNow that is used in a multinational healthcare company. The data export is based on some filters such as CI and the contact point of incidents. In order to make sure that only English data is exported, a filter for the contact point was applied and only the incidents that have been reported to the international support centers, which provide support in English, were selected; the incidents that have been reported to local support centers with local languages were eliminated.

In order to compare the impact of the dataset and the number of classes, the tests were run on 2 different datasets. First dataset, Dataset A, contains 34,006 documents with 8 classes while the second dataset, Dataset B, contains 96,586 documents with 52 classes. The classes are the CI’s and the top classes were filtered when exporting the incidents from ServiceNow. Dataset A contains incidents with top 8 CI’s based on the amount of incidents created using the CI and Dataset B contains incidents with top 52 CI’s. In Dataset B, the filter was also extended to some specific English speaking countries such as United Kingdom and United States of America in order to get more results in English language and expand the dataset.

While progressing with the experiments, a class with low TP’s was identified and this raised the necessity to create a new dataset, Dataset C, by removing that specific class with the low TP rate from the dataset. Table 5.1 lists the number of documents and classes of the datasets.

Table 5.1 Dataset details

	Number of Incidents	Number of Classes
Dataset A	34,006	8
Dataset B	96,586	52
Dataset C	28,215	7

Although the datasets are not balanced, we avoided the scaling of the dataset in order to work with the original data. For privacy purposes, some alterations were made on the data and class names have been changed as much as possible with the maximum effort on renaming them in the way to keep the main idea of their original names. In Table 5.2, Table 5.3 and Table 5.4, the number of documents in the datasets are listed.

Table 5.2 Dataset A class distribution

Class Name	Number of Documents
Accounts & Access	5,791
E-mail Server	7,815
Folder Access	5,466
Mobile Application 1	3,770
ERP Customer Experience BusinessUnit1	2,203
ERP - Order Fulfillment	2,853
ERP Security	5,294
Learning Management System	814

Table 5.3 Dataset B class distribution

Class Name	Number of Documents
Servicenow-It	1,793
E-mail Server	11,303
Printer-Network	982
Learning Management System	1,627
Folder Access	7,590
Microsoft Office 2016	847
Laptop	2,805
Active Directory - Domain1	3,735
ERP Security	8,064
Regulatory Application	1,322
ERP portal 1	1,688
Mobile Application 1	4,115
ERP - Order Fulfillment	3,783
Accounts & Access - Cross Company	7,449
CRM-Italy	1,653
Microsoft Outlook	1,470
Sales Analyzer Application 1	1,039
Accounts & Access	777
Material Management Application	1,099
Customer Sales Reporting	3,017
Internal application 1	947
Hardware	802
ERP Customer Experience	2,408
IT Portal	925
ServiceNow	1,377

Table 5.3 Dataset B class distribution (cont.)

ERP Client	1,262
Compliance Application	1,823
Improvements	747
Customer Sales Reporting Access	763
iPhone	808
Printer Hardware	794
Team sites	758
ERP Sales Management	1,673
Invoicing	835
Identity Manager	742
Online Meeting	776
Network	1,155
ERP - Supplier Management	1,657
Computer New/Reimage	811
Legacy Learning Management System	1,188
Outlook 2010 Hotfix	3
IP Phones	1,358
Desktop/Laptop/Mobile Computing Management	1,152
Business Intelligence Service	2,035
Microsoft Outlook-Domain2	9
Secure File Sharing	3
ERP Non-production	1,519
Shipping	913
Encryption	2
Internal application 2	1,176
Outlook plug-in for Secure File Sharing	6
Outlook addon for online meeting	1

Table 5.4 Dataset C class distribution

Class Name	Number of Documents
E-mail Server	7,815
Folder Access	5,466
Mobile Application 1	3,770
ERP Customer Experience	2,203
ERP - Order Fulfillment	2,853
ERP Security	5,294
Learning Management System	814

5.2 Pre-processing

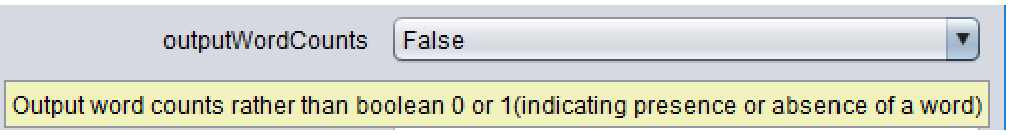
WEKA, an open source tool developed in Waikato University, is used for the text classification tasks in this study. In order to use the tickets in Weka, the data extracted from ServiceNow had to be converted to .arff format in order to load in Weka. After loading the incidents in Weka, the text strings in the incidents were transformed into a Term Frequency x Inverse Document Frequency (TF-IDF) vector using the pre-processing function “StringToWordVector”. This way a vector was formed with the weight each term in each document. A weight of a term “i” in document “j” can be formulated as in (5.1) where $tf_{i,j}$ is the number of occurences of term “i” in document “j”, N is the total number of documents and df_i is the number of documents containing “i”. (Salton & Buckley, 1988)

$$w_{i,j} = tf_{i,j} \times \left(\log \frac{N}{df_i}\right) \qquad (5.1)$$

In order to achieve the highest possible accuracy, some pre-processing methods were applied on the datasets. The first pre-processing step was removing the stopwords, the common words such as “the”, “a”, “are” which have no impact on the meaning of a document, using the “Rainbow” stopword list by adding the company name to the stopword list. The incidents in the dataset were received via e-mail messages and most of the users have their e-signatures, which contains the company name, in their e-mail messages. After this all terms, or words, were converted to lowercase in order to prevent considering the same words as different terms. Terms “RE:” and “FW” were also removed as they are sometimes part of the e-mail subject in case of replies of forwards.

Combinations of two other pre-processing steps were applied as an argument to compare and comprehend their impact on the accuracy. These two steps are outputWordCount function and alphabetic conversion.

With outputWordCounts true, the weight of a term in the TF-IDF vector is impacted by the number of occurrences of the term in a document. An explanation of this option is also described in Figure 5.1. When outputWordCounts is false, the number of occurrences of the term has no impact on the term weight. Instead, the weight then depends on the existence of the term; not the number of occurrences.



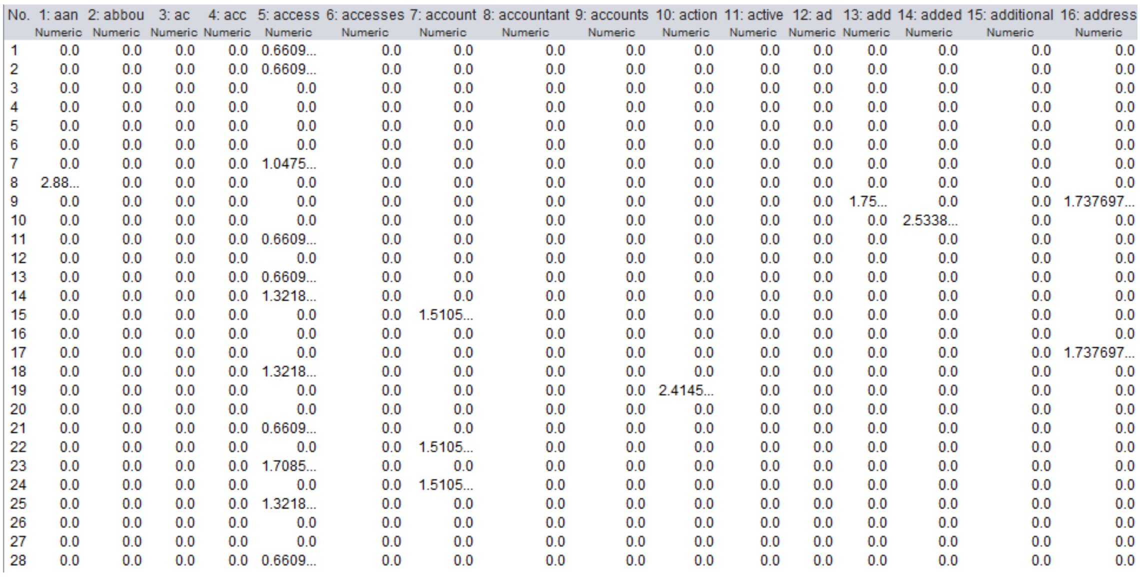
outputWordCounts: False

Output word counts rather than boolean 0 or 1(indicating presence or absence of a word)

Figure 5.1 outputWordCounts description in Weka

Alphabetic conversion removes the terms without non-alphabetic characters. By using this parameter the impact of the terms with special characters or numbers on the accuracy was measured.

A sample from the TF-IDF vector produced after applying alphabetic and word count filters can be seen in Figure 5.2.



No.	1: aan	2: abbou	3: ac	4: acc	5: access	6: accesses	7: account	8: accountant	9: accounts	10: action	11: active	12: ad	13: add	14: added	15: additional	16: address
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	0.0	0.0	0.0	0.0	0.6609...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.6609...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	1.0475...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	2.88...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.75...	0.0	0.0	1.737697...
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.5338...	0.0	0.0
11	0.0	0.0	0.0	0.0	0.6609...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	0.0	0.0	0.0	0.0	0.6609...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	0.0	0.0	0.0	0.0	1.3218...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	0.0	0.0	0.0	0.0	0.0	0.0	1.5105...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.737697...
18	0.0	0.0	0.0	0.0	1.3218...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.4145...	0.0	0.0	0.0	0.0	0.0	0.0
20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
21	0.0	0.0	0.0	0.0	0.6609...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	0.0	0.0	0.0	0.0	0.0	0.0	1.5105...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
23	0.0	0.0	0.0	0.0	1.7085...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
24	0.0	0.0	0.0	0.0	0.0	0.0	1.5105...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25	0.0	0.0	0.0	0.0	1.3218...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
28	0.0	0.0	0.0	0.0	0.6609...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 5.2 Sample from TF-IDF vector

5.3 Test Results

Holdout method was used on splitting datasets into train and test splits. Our dataset was split in two parts: First partition, which contains 80% of the whole dataset, is used as the training data and the remaining 20% used as the test data. This rate was decided based on the related work by Zinner et.al (2015). Due to memory limitations, it was not possible to use the n-fold cross validation method.

The results of the application of Naïve Bayes Multinomial, SVM, C4.5 and kNN (k=7) on Dataset A are listed below in Table 5.5.

Table 5.5 Results from Dataset A

Algorithm	Characters	OutputWordCount	Accuracy (%)	Time to Train (s)	Time to Test (s)
C4.5	All Characters	TRUE	72.81	4,880	0.62
		FALSE	72.64	7,801	1.07
	Alphabetic Only	TRUE	72.49	5,312	0.6
		FALSE	73.15	10,048	0.98
kNN	All Characters	TRUE	50.80	0.05	152.22
		FALSE	49.63	**0.02**	**152.74**
	Alphabetic Only	TRUE	59.22	**0.02**	123.84
		FALSE	49.34	0.04	152.37
Naïve Bayes Multinomial	All Characters	TRUE	69.20	0.17	0.99
		FALSE	68.50	0.18	1.07
	Alphabetic Only	TRUE	69.93	0.13	1.06
		FALSE	69.92	0.12	0.85
SVM	All Characters	TRUE	78.91	2,390	2.95
		FALSE	78.62	2,569	2.72
	Alphabetic Only	TRUE	**79.71**	1,883	2.9
		FALSE	79.47	2,130	2.94

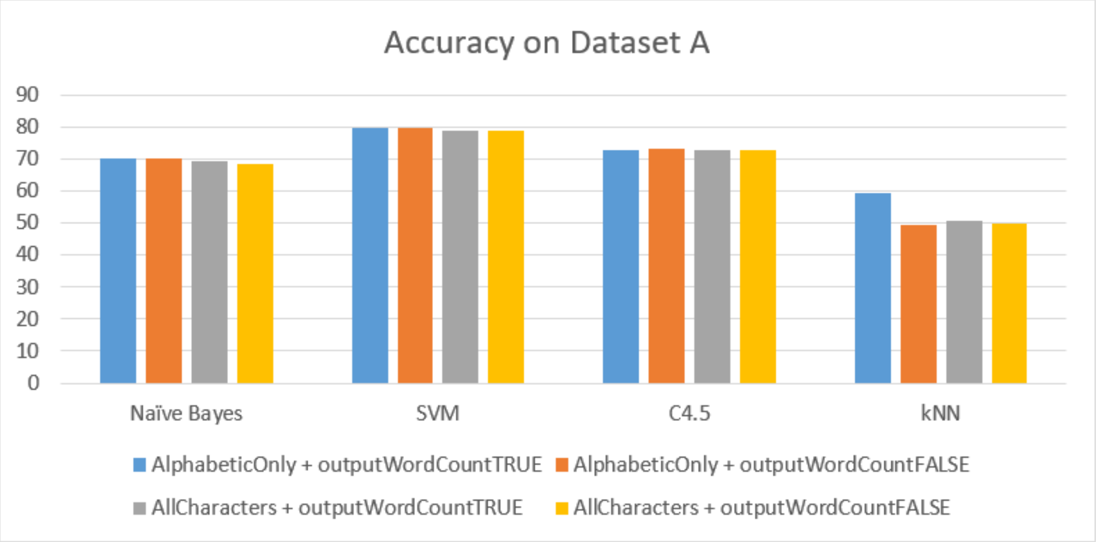


Figure 5.3 Accuracy on Dataset A

The experiments, as shown in Table 5.5 and Figure 5.3, on Dataset A proved that SVM performs the highest accuracy on identifying the class on the test data by reaching almost 80% accuracy on predictions no matter what the pre-processing step is. SVM is followed by C4.5 with over 70%, Naïve Bayes Multinomial with almost 70% and kNN with below 60% respectively.

Obviously SVM outperforms the other algorithms on the accuracy in this experiment but the long time required for training the model in SVM might lead us to choosing performance over accuracy. Naïve Bayes Multinomial has around 10% lower performance on accuracy but the training time required for Naïve Bayes Multinomial is very low compared to SVM. As seen In Figure 5.4, the time required for training the model for Naïve Bayes Multinomial and kNN are almost zero.

Another point to take note on the results is that alphabetic terms with word occurrences gave the best accuracy on all algorithms except for C4.5. This is also the case on training time, the training time is relatively low when word occurrences and only alphabetic characters are enabled.

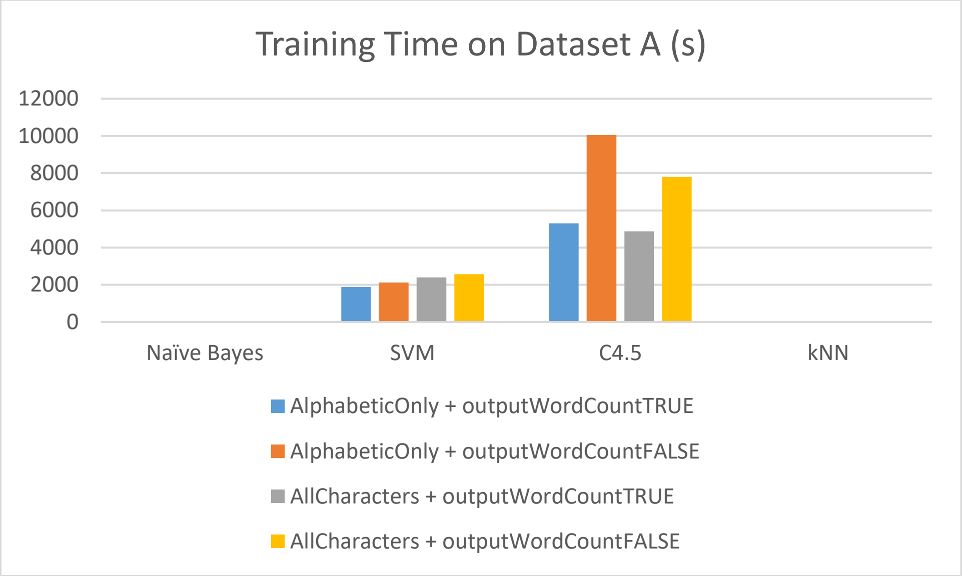


Figure 5.4 Training Time on Dataset A

In case of a dilemma between accuracy and training time, it's strongly recommended to first optimize the dataset or operations for the classification before making a decision on what algorithm to use. The analysis on the results of Dataset A showed that the TP rate of class "Accounts & Access" was low for all experiments. An example of this can be seen in Table 5.6. The reason behind this is that the access requests for ERP, folders and E-mail accounts are similar. It is very important to emphasize the difference of these CI's in the configuration management database in order to prevent confusions on classifying the incidents.

This led us removing the class "Accounts & Access" from Dataset A creating a new dataset with the name of Dataset C in order to measure the impact of this single class. From business or operation perspective, a suggest the business owner would be to optimize and review the CI's stored in the CI database to make sure that different CI's for the same purpose do not exist or the difference between CI's are clear.

Table 5.6 Analysis on Naïve Bayes Multinomial results applied on Dataset A.

TP Rate	FP Rate	Precision	Recall	Class
0.354	0.068	0.521	0.354	Accounts & Access
0.750	0.066	0.769	0.750	E-mail Server
0.808	0.051	0.751	0.808	Folder Access
0.877	0.033	0.772	0.877	Mobile Application1
0.939	0.012	0.846	0.939	ERP Customer Experience BusinessUnit1
0.761	0.044	0.614	0.761	ERP-OrderFulfillment
0.620	0.055	0.671	0.620	ERP Security
0.774	0.024	0.449	0.774	Learning Management System

In Dataset B, the experiments took longer, compared to Dataset A, due to the larger amount of documents and classes. It took over 3 days for C4.5 decision tree for training. The accuracy, also, on Dataset B was lower due to the complexity of the dataset. SVM is again the top performer based on accuracy with 64.41% correct classifications. Although the training time for SVM is again high compared to other algorithms, this training time can be tolerated. C4.5 made it close to SVM with 57.60% correct classifications but it is a challenge to overcome the long training time for C4.5 and this makes SVM more effective than C4.5. Table 5.7 and Figure 5.5 show how the algorithms performed on Dataset B.

Table 5.7 Results from Dataset B

Algorithm	Characters	OutputWordCount	Accuracy (%)	Time to Train (s)	Time to Test (s)
C4.5	All Characters	TRUE	57.04	330,070.47	16.53
		FALSE	56.89	321,800.15	13.85
	Alphabetic Only	TRUE	57.38	277,458.29	24.48
		FALSE	57.38	262,052.34	11.66
kNN	All Characters	TRUE	36.50	0.09	1,706.98
		FALSE	34.51	**0.08**	1,631.97
	Alphabetic Only	TRUE	36.83	0.09	1,756.15
		FALSE	35.14	0.11	1,703.63
Naïve Bayes Multinomial	All Characters	TRUE	48.93	0.27	4.61
		FALSE	47.11	0.29	5.72
	Alphabetic Only	TRUE	49.85	0.28	**3.75**
		FALSE	48.56	0.26	5.87
SVM	All Characters	TRUE	63.86	10,720.33	230.02
		FALSE	62.65	11,847.48	233.17
	Alphabetic Only	TRUE	**64.41**	9,707.32	229.58
		FALSE	63.36	9,964.05	228.3

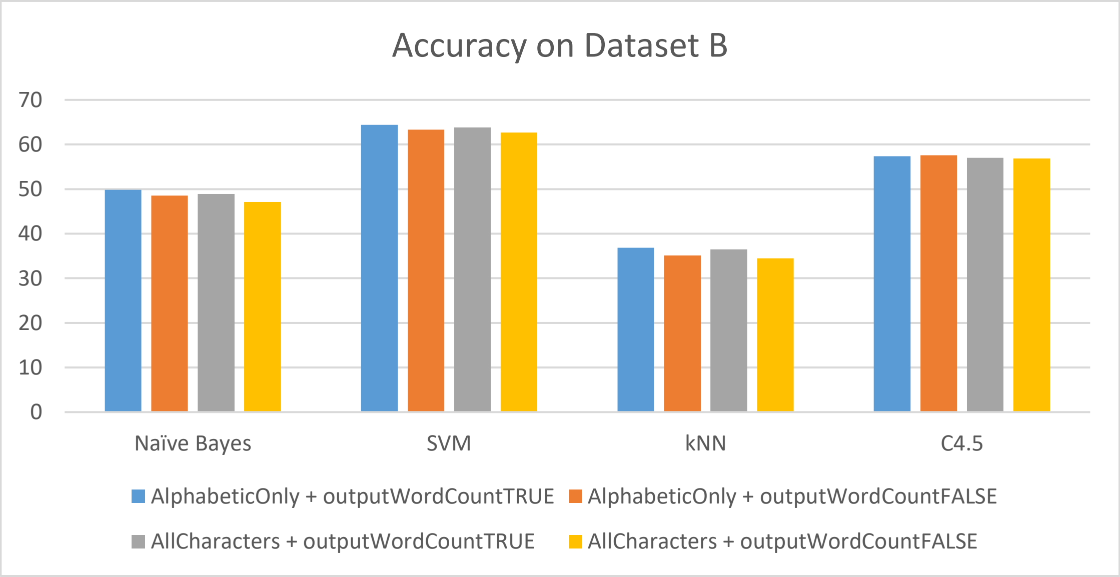


Figure 5.5 Accuracy on Dataset B

From accuracy perspective, the pre-processing parameters did as small impact as on Dataset A but it’s still possible to consider alphabetic only terms with occurrences (outputWordCount=TRUE) gave the best accuracy results even though it’s a small rate that made the improvement.

Training time was a challenge in this large dataset with 50+ classes, especially with C4.5 algorithm. It took over 3 days to successfully complete a single C4.5 algorithm on this dataset. Apart from this, training time ratio for other three algorithms were close to Dataset A. Naïve Bayes Multinomial and kNN took milliseconds to train and SVM took roughly 10000 seconds to train as shown in Figure 5.6.

A quick comparison between the training times and accuracies leads to the conclusion that SVM would be the best algorithm for a complex dataset like Dataset B. Although the maximum accuracy achieved is 64.41%, which is not perfect, it can be useful for nice to have functions like auto-suggestion of CI in the application.

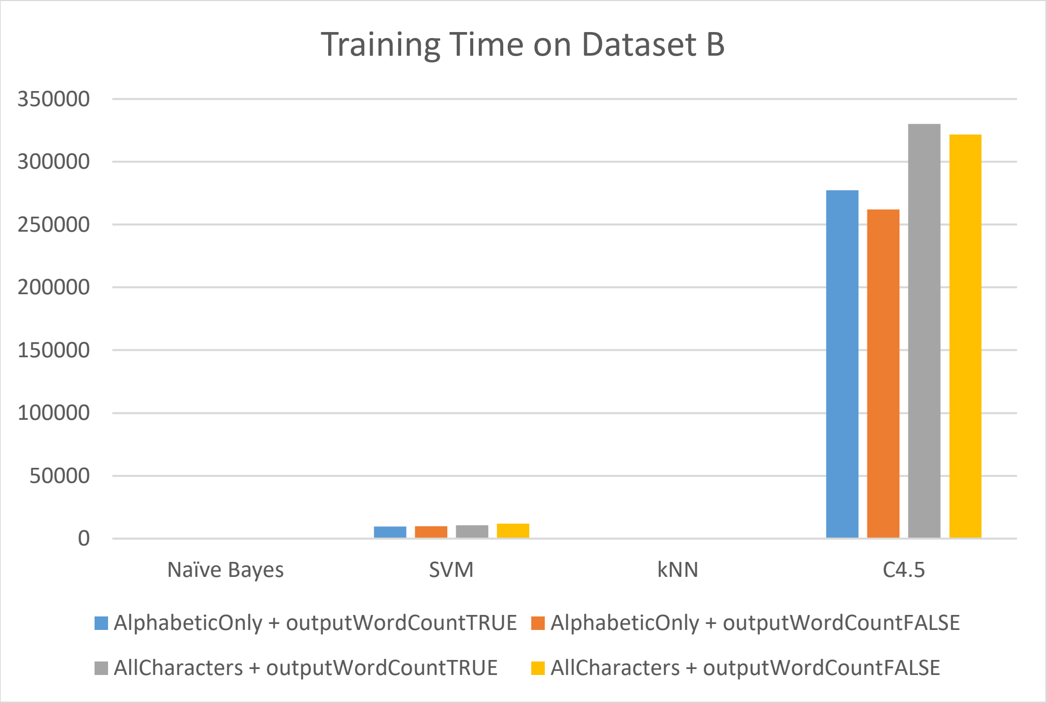


Figure 5.6 Training Time on Dataset B

Based on the results on Dataset A, we formed a new dataset with the name of Dataset C by excluding the documents with class “Accounts & Access” from Dataset A because of the low TP rate of this specific class. The reason behind this is the similarity between “Accounts & Access” and other two classes “E-mail Server” and “Folder Access”. Results for this new dataset are displayed in Table 5.8 and Figure 5.7.

Table 5.8 Results from Dataset C

Algorithm	Alphabetic	OutputWordCount	Accuracy (%)	Time to Train (s)	Time to Test (s)
C4.5	All Characters	TRUE	83.77	4,860.87	0.74
		FALSE	83.71	5,531.71	0.58
	Alphabetic Only	TRUE	84.53	3,233.91	0.38
		FALSE	84.55	4,078.66	0.56
kNN	All Characters	TRUE	59.63	0.03	133.17
		FALSE	59.44	0.03	130.88
	Alphabetic Only	TRUE	60.22	0.03	116.61
		FALSE	59.19	**0.03**	125.72
Naïve Bayes Multinomial	All Characters	TRUE	79.73	0.11	0.36
		FALSE	79.09	0.11	0.32
	Alphabetic Only	TRUE	80.52	0.12	0.59
		FALSE	80.15	0.11	**0.4**
SVM	All Characters	TRUE	89.12	586.02	1.16
		FALSE	89.03	643.24	1.11
	Alphabetic Only	TRUE	**89.72**	472.94	1.12
		FALSE	89.56	613.89	1.15

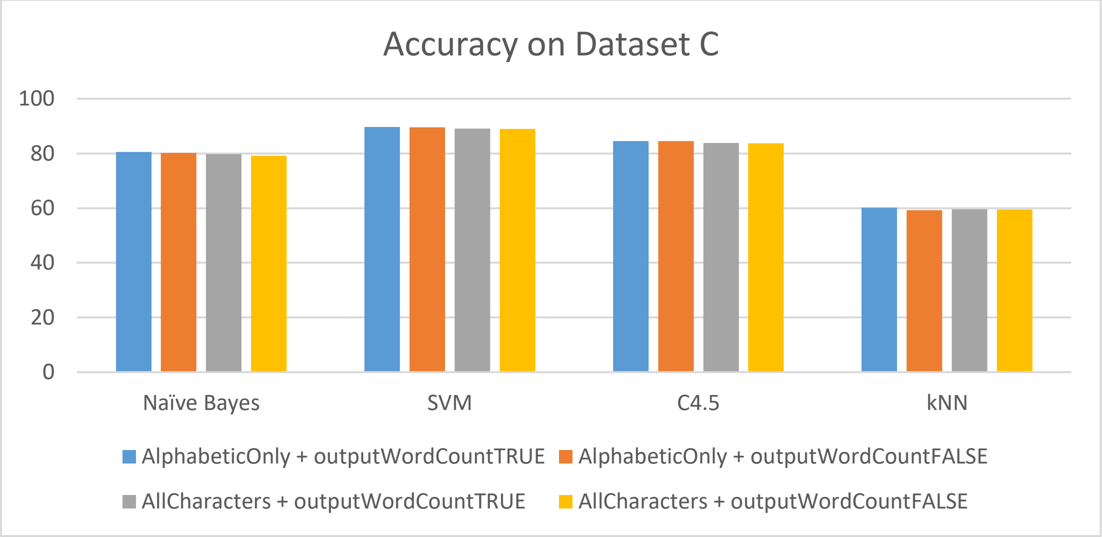


Figure 5.7 Accuracy on Dataset C

As seen in Figure 5.7, simply by removing a class from the dataset, it was possible to reach 80% accuracy on three algorithms while kNN performed the lowest accuracy which is below 60% on Dataset C. Top performer on accuracy again is SVM with the highest score of 89.72%. Just like in other datasets, number of word occurrences and removing terms with non-alphabetical terms improved the accuracy even if it’s a very small rate.

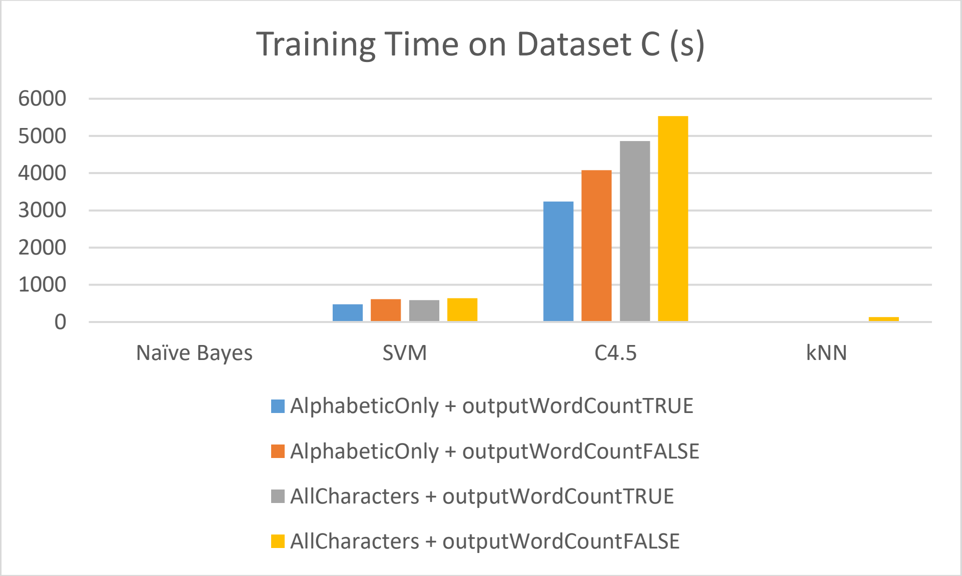


Figure 5.8 Training time on Dataset C

From training time perspective shown in Figure 5.8, kNN and Naïve Bayes took the shortest time for training and C4.5 takes quite long compared to the other three algorithms. SVM's training time dropped by over 70% compared to Dataset A and this was done by simply removing 5791 documents out of 34006. C4.5 also performed well on accuracy but the high training time for it is a reason for not prioritizing this algorithm in this area of study. Considering the low training time and relatively high accuracy, even though it's not as high as C4.5 and SVM, of Naïve Bayes Multinomial makes it a good choice with this dataset.

6. CONCLUSION

In this study, an automatic solution was suggested to the problem of manual categorization of IT incidents, based on their CI's, in ITSM softwares that are based on ITIL framework by using text classification methods. This task causes loss of seconds for each incident in daily operations and it could be summed up as minutes or even hours every week when it's about thousands of incidents to be categorized.

Various combinations of classification algorithms and pre-processing methods on 3 different datasets were used in the experiments and different levels of accuracies were achieved. In all experiments performed in this study, SVM proved to achieve the highest accuracy but the training time was not very low for it. It was obvious that C4.5 is a very costly classifier from training perspective compared to the other three algorithms. Naïve Bayes Multinomial and kNN have very low training time due to their simple probabilistic and distance based approaches. However, as the dataset becomes more complex, accuracy of Naïve Bayes Multinomial and kNN drops dramatically while the accuracy of C4.5 and SVM drop in a tolerable way.

From pre-processing perspective, different combinations based on word occurrences and the character limitations were applied and compared for their impact and the best accuracy was achieved in the scenarios where the terms were weighted based on the number of their occurrences (outputWordCount=TRUE) and with the exclusion of non-alphabetic characters from the dataset. This pre-processing approach also helped on reducing the time spent on training the model.

Apart from the technical measurements of classification methods, it was also found out that a good optimization needs to be applied on the dataset before starting the evaluation. Comparison of Dataset A and Dataset C is the evidence that having a class that is similar to another reduces the accuracy in a considerable rate. The accuracy was improved by 10% by removing a single class which is about 17% of the dataset in size. The main reason for the improvement was the similarity between this specific class and to another. It's even difficult for a person to distinguish when a request related to e-mail access is received: Is it a case of Accounts & Access or E-mail Server? If an automated classification is required for such problems, the classes should be as distinctive as possible for the best accuracy.

This study can be enhanced by using the ensemble methods, which is the technique of using a combination of different algorithms with various voting approaches, as well as other classifiers.

Another possible enhancement to this study would be using a dataset with a combination of languages. It’s possible to extract datasets (incidents) from the same database with a combination of different languages and this would enable a multi-label approach as the instances will belong to two different classes; one from language perspective and the other one from a CI perspective.

REFERENCES

Agarwal, S., Aggarwal, V., Akula, A.R., Dasgupta, G.B., Sridhara, G. (2017) Automatic problem extraction and analysis from unstructured text in it tickets. *IBM J. Res. Dev. 61, pp. 4–41.*

Agarwal, S., Sindhgatta, R., Sengupta, B. (2012) SmartDispatch: enabling efficient ticket dispatch in an it service environment. *In: 18th ACM SIGKDD 2012.*

Altintas, M., Tantug, A.C. (2014) Machine Learning Based Ticket Classification in Issue Tracking Systems.

Arcilla, M., Calvo-Manzano, J. A., San Feliu, T. (2013). Building an IT service catalog in a small company as the main input for the IT financial management. *Computer Standards & Interfaces, 36(1), pp. 42–53*

Beneker, D., Gips, C. (2017). Using Clustering for Categorization of Support Tickets. *LWDA.*

Cartlidge, A., Rudd, C., Smith, M., Wigzel, P., Rance, S., Shaw, S., Wright, T. (2012). An Introductory Overview of ITIL® 2011. London: The Stationery Office

Comptia (2017). https://certification.comptia.org/it-career-news/post/view/2017/08/30/itsm-frameworks-explained-which-are-most-popular Access date: May 23, 2019

Frank E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I., Trigg, L. (2009). Weka-A Machine Learning Workbench for Data Mining. *In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA, pp. 1269-1277.*

Gehrmann, M. (2012). Combining ITIL, COBIT and ISO/IEC 27002 for structuring comprehensive information technology for management in organizations. *NAVUS: Revista de Gestão e Tecnologia, pp. 66-77*

Gil-Gomez H., Oltra-Badenes, R., Adarme-Jaimes, W. (2014) Service Quality Management Based on the Application of the ITIL Standard. *Dyna v.81, n.186, pp.51-56.*

Han, J., Kamber, M., Pei, J. (2011) *Data Mining: Concepts and Techniques (3rd edition)*. San Francisco: Morgan Kaufmann.

Joachims, T. (1998) Text categorization with Support Vector Machines: learning with many relevant features. *In Proceedings of the 10th European Conference on Machine Learning (ECML'98), Claire Nédellec and Céline Rouveirol (Eds.). Springer-Verlag, Berlin, Heidelberg, pp. 137-142.*

McCallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification.

Orand, B., Villarreal, J. (2011). *Foundations of IT Service Management.*

Pratama, B., Sarno, R (2015) Personality Classification Based on Twitter Text Using Naive Bayes, KNN and SVM. *International Conference on Data and Software Engineering (ICoDSE), pp. 170–174,*

Qin, Y., Wang, X. (2009). Study on Multi-label Text Classification Based on SVM. *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery.*

Sakolnakorn, P.P., Meesad, P., Clayton, G. (2008). Automatic Resolver Group Assignment of IT Service Desk Outsourcing in Banking Business.

Salton, G., Buckley, C. (1988) Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage. 24, 5, 513-523.*

Şeker, Ş. (2013) *İş Zekası ve Veri Madenciliği.* Cinius Yayınları, Istanbul.

Zinner T., Lemmerich F., Schwarzmann S., Hirth M., Karg P., Hotho A. (2015) Text Categorization for Deriving the Application Quality in Enterprises Using Ticketing Systems. *In: Madria S., Hara T. (eds) Big Data Analytics and Knowledge Discovery. DaWaK 2015. Lecture Notes in Computer Science, vol 9263. Springer, Cham.*

RESUME

Personal Information

Name	: Erdal Sever İşcen
Date of Birth	: 05.07.1986
E-mail	: severiscen@gmail.com
Phone	: +90 530 320 6282

Education

Graduate	: Computer and Information Systems at Doğuş University (2008 - present)
Bachelor	: Computer Engineering at Doğuş University (2004 - 2008)
High School	: Kartal Burak Bora Anatolian High School (2003)

Work Experience

- Medtronic Medikal Teknoloji Tic. Ltd. Şti.

  Sr. Technical User Support Analyst – Healthcare IT (2016 – present)

  IT Technologist – Local IT (2009 – 2016)