İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

KULLANICI DESTEK SİSTEMLERİNDE YARDIM BİLETLERİNİN OTOMATİK SINIFLANDIRILMASI

YÜKSEK LİSANS TEZİ

Mücahit ALTINTAŞ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Eylül 2014

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

KULLANICI DESTEK SİSTEMLERİNDE YARDIM BİLETLERİNİN OTOMATİK SINIFLANDIRILMASI

YÜKSEK LİSANS TEZİ

Mücahit ALTINTAŞ
(504111545)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Yrd. Doç. Dr. A. Cüneyd TANTUĞ

Eylül 2014

İTÜ, Fen Bilimleri Enstitüsü'nün 504111545 numaralı Yüksek Lisans Öğrencisi Mücahit ALTINTAŞ, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "KULLANICI DESTEK SİSTEMLERİNDE YARDIM BİLETLERİNİN OTOMATİK SINIFLANDIRILMASI" başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı : **Yrd. Doç. Dr. A. Cüneyd TANTUĞ** ……..
İstanbul Teknik Üniversitesi

Jüri Üyeleri : **Yrd. Doç. Dr. Tolga OVATMAN** ……..
İstanbul Teknik Üniversitesi

Yrd. Doç. Dr. Arzucan ÖZGÜR ……..
Boğaziçi Üniversitesi

Teslim Tarihi : 25 Ağustos 2014
Savunma Tarihi : 2 Eylül 2014

Aileme ve tüm sevdiklerime,

ÖNSÖZ

Son yıllarda artan mobil cihaz kullanımı ve internet erişim olanağı, insanların cihazlarla olan etkileşim süresini arttırmıştır. Dolayısıyla insan ve cihazlar arasındaki etkileşimin mümkün olan en uygun hale getirilmesi, her zamankinden daha gerekli hale gelmiştir. Bu etkileşimin en uygun şekilde gerçekleşmesi için, çeşitli yöntem ve araçlar geliştirilmiştir. Bu yöntem ve araçlar geliştirilirken, insanlara rahat kullanım imkânı sağlamanın yanısıra en verimli sonucu alabilmek hedeflenmiştir. Bu sebeble son yıllarda, cihazlara ve sistemlere doğal dili yorumlayabilme yeteneği kazandırılmaya çalışılmaktadır. Bu çalışmada, yapay zeka teknikleri ve doğal dil işleme yöntemleri kullanılarak doğal dil ile yazılmış düzensiz formdaki bir yardım biletinin değerlendirilip ilgili kişiye aktarılması ve hatta cevaplanması amaçlanmıştır. Konu üzerinde araştırma yapanlara faydalı olmasını temenni ederim.

Bu çalışmada İTÜ Destek Sistemi verileri kullanılmıştır. Bu verileri paylaştıkları için başta Bilgi İşlem Daire Başkanı ve aynı zamanda danışmanım olan Cüneyd TANTUĞ hocama ve bu konuda yardımcı olan diğer arkadaşlara teşekkür ederim.

Ayrıca, bugüne kadar benden maddi ve manevi desteklerini hiçbir zaman esirgemeyen aileme, bilgisi ve tecrübesiyle beni aydınlatan değerli hocalarıma sonsuz teşekkürlerimi sunarım.

Eylül 2014

Mücahit ALTINTAŞ
(Bilgisayar Mühendisi)

İÇİNDEKİLER

Sayfa

KISALTMALAR

DDİ : Doğal Dil İşleme
idf : inverse document frequency
ITIL : Information Technology Infrastructure Library
kNN : k-Nearest Neighbors
ML : Machine Learning
NLP : Natural Language Processing
SSS : Sık Sorulan Sorular
SVM : Support Vector Machine
tf : term frequency
tf-idf : term frequency–inverse document frequency

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

KULLANICI DESTEK SİSTEMLERİNDE YARDIM BİLETLERİNİN OTOMATİK SINIFLANDIRILMASI

ÖZET

Günümüz rekabet koşullarında, kullanıcı destek sistemlerinin sağladığı servis kalitesi, ürün veya hizmet alımı öncesi ve sonrasında müşterilerin organizasyon hakkındaki memnuniyetini belirleyen önemli bir etken haline gelmiştir. Organizasyonlar müşterilerinin veya kullanıcılarının sorun, görüş ve isteklerini en etkin şekilde değerlendirip cevaplamak suretiyle müşteri memnuniyetini arttırmayı amaçlamaktadırlar. Kullanıcı destek sistemleri kullanıcıdan gelen sorun, görüş ve istekleri yardım bileti şeklinde tutup, cevaplanmak üzere ilgili destek personeline veya birimine atanmasını sağlarlar. Günümüzdeki destek sistemlerinin bir kısmı bu atama işlemini, daha önceden belirlenmiş seçenekleri kullanıcıya işaretleterek yaparken, diğer bir kısmı operatör veya insan sınıflandırıcı vasıtasıyla gerçekleştirmektedir. Her iki yöntem de zaman alıcı ve hataya meyilli insan çabası gerektirir. Bu durum, özellikle büyük kuruluşlarda, kaynakların verimsiz şekilde kullanılmasına neden olmakta ve kullanıcı memnuniyetini olumsuz yönde etkilemektedir. Bu çalışmada kullanıcı destek sistemlerinin kalitesini iyileştirmek amacıyla, yeni bir destek sistem mimarisi önerilmiştir. Yapay zekâ algoritmaları ve doğal dil işleme yöntemleri kullanılarak yardım biletlerinin ilgili birime otomatik olarak atanması gerçekleştirilmiş, bunun yanı sıra tekrar eden yardım biletlerinin sistem tarafından otomatik cevaplanması sağlanmıştır. Metin halindeki verinin sayısal hale dönüştürülmesi için *bag of word* yaklaşımından yola çıkarak, veri kümesi içinde bulunan her terim birbirinden bağımsız nitelikler şeklinde kabul edilmiştir. Bu yaklaşımdan dolayı metin sınıflandırmada kullanılan özellik vektörü boyutu hayli büyük olmaktadır. Özellikle Türkçe, Japonca, Macarca gibi eklemeli dillerde bir terim pek çok farklı formda bulunabilmektedir. Bu durum ayrıca veri seyrekliği problemini de doğurmaktadır. Bu yüzden veri kümesi sırasıyla biçimbilimsel analiz ve biçimbilimsel belirsizlik giderimi işlemlerine tabi tutulmuştur. Sınıflandırma için ayırt edici bir özelliğe sahip olmayan düşük idf değerine sahip terimler, konudan bağımsız olarak kullanılan bağlaçlar ve sıradan kelimeler (stop words) özellik vektöründen çıkarılmıştır. Bu sayede mümkün olduğunca, özellik vektör boyutu küçültülmeye çalışılmış ve özellik vektörü gürültüden arındırılmıştır. Ayrıca kullanılan veri kümesindeki yazım yanlışları ve sayısal ifadelerden kurtulmak için ön veri aşamasında bir takım işlemler gerçekleştirilmiştir. Önerilen sistem sayesinde; yardım biletlerinin ilgili birime atanma işlemi otomatikleştirilerek bu işlemin destek personeli üzerindeki yükü, azaltılmıştır. Bu sayede zaman alıcı ve hataya meyilli insan emeği gerektiren elle atama işlemi asgari düzeye indirgenmiştir. Atama işleminin otomatikleştirilmesiyle, kullanıcıdan gelen yardım biletleri daha kısa sürede cevaplanarak sağlanan servis kalitesi, dolayısıyla son-kullanıcı memnuniyeti arttırılmıştır. Yanlış atama işlemleri minimuma indirgenerek, değerli destek kaynaklarının daha etkin ve verimli şekilde kullanılmasına olanak sağlanmıştır. Ayrıca, tekrar eden yardım biletlerinin tespit edilip otomatik cevaplandırılması, destek personelinin iş yükünü azaltmıştır.

AUTOMATIC CLASSIFICATION OF HELP TICKETS IN USER SUPPORT SYSTEMS

SUMMARY

In the current competitive environment, the service quality of user support systems, before and after purchase of products or services, has become an important factor that affect customer satisfaction. A user suppport system or Issue Tracking System (ITS) is a type of software to capture and keep track of customer issues, which may be customer problems or customer requests, and to assign the tickets the relevant support person or unit in order to create a solution. 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. 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 recourses. 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 relies 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 enduser satisfaction.

In this study, an extension to an ITS for automatically assigning the issue tickets to the right related 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. Additionaly, in this study, automatic response of repetive tickets is proposed,. In most case, users ignore to look previously asked questions and ask the similar questions. The proposed system intends to use resources in the most efficient manner. So, help tickets are firstly tried to be answered without any assistance of technician expert. The similarity of each FAQ between coming help ticket is calculated. The answer of the most similar FAQ which exceeds predefined threshold similarity is suggested to the user. If the user does not become satisfied with the suggestion, the ticket is escalated the up level to be responded by technical person.

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. 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.

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. 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.

Not distinctive terms for classification which have got smaller 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.

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

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.

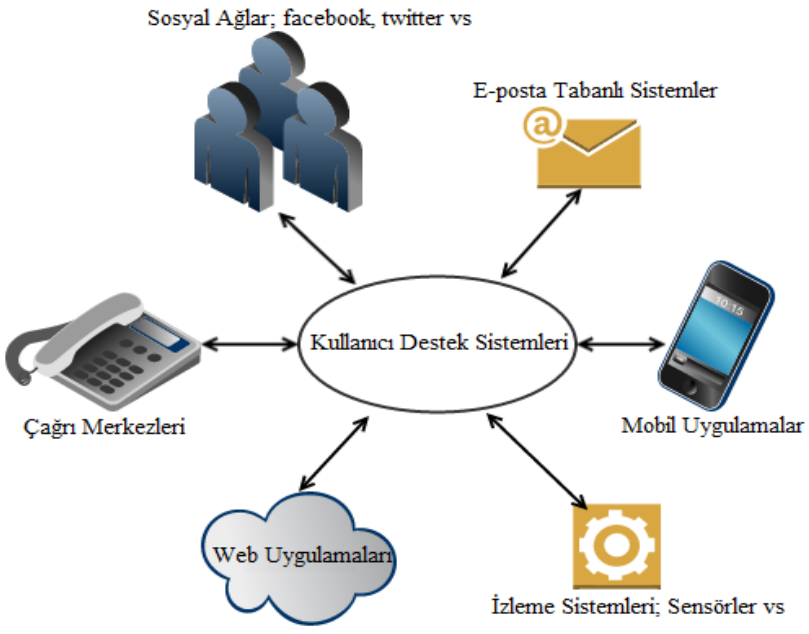
Nearest neighbor algorithm aims to predict label of a new instance by measuring distance of the closest predefined number of training samples. This algorithm is commonly preferred when there is little or no prior knowledge about the distribution of the data. It is an instance based classification algorithm.

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.

Briefly, 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 large organizations better allocation and effective usage of the valuable support resources.

1. GİRİŞ

Günümüz rekabet koşulları, organizasyon ve hedef kitle arasındaki olumlu ilişkileri üstünlük sağlama adına önemli bir etken haline getirmiştir. Teknolojik gelişmeler ve diğer uygulamalar girişimcileri işletme kurmaya teşvik etmiş, bunun neticesi olarak aynı kalitede mal üreten veya hizmet veren çok sayıda firma ortaya çıkmıştır. Böyle bir ortamda müşteri hizmetleri önem kazanmış ve ciddi bir rekabet unsuru haline gelmiştir. Organizasyonlar müşteriler ile olumlu ilişkiler oluşturmak amacıyla yeni strateji ve sistemler geliştirmişlerdir. Geliştirilen bu sistemlerden biri de sorun takip sistemleri veya yardım masası olarak da isimlendirilen müşteri destek sistemleridir. Bugün, müşteri destek sistemlerinin kalitesi müşterilerin tercihini belirlemede önemli rol oynamaktadır. Müşteri destek sistemleri, hizmet veya ürün alımı öncesi ve sonrasındaki süreçlerde, müşteri memnuniyetini arttırmak üzere oluşturulmuş yapılardır. Müşteri destek sistemlerinin kullanıcıları, organizasyonun müşterileri olabilecekleri gibi, organizasyon tarafından sağlanan servisin son kullanıcıları da olabilirler. Bu yüzden bu çalışmada müşteri ve kullanıcı ifadeleri birbirlerinin yerine kullanılmıştır.



Şekil 1.1: Kullanıcı destek sistemlerinin akış kanalları

Bu sistemler çeşitli kanallar üzerinden gelen görüş ve istekleri toplayıp, destek birimi içerisindeki uygun teknik personele aktararak, gelen görüş ve isteklerin mümkün olan en kısa sürede değerlendirilip yanıtlanmasını amaçlarlar (Şekil 1.1). Kullanıcılar destek sistemlerine; çağrı merkezleri, web ve mobil uygulamalar, e-posta tabanlı sistemler aracılığıyla ulaşabilirler. Bunların yanı sıra cihazların durumunu izleyen algılayıcılar (sensörler) vs. vasıtasıyla kullanıcı destek sistemlerine bilgi akışı sağlanabilir. Son yıllarda organizasyonlar, tüm bunlara ek olarak facebook ve twitter gibi sosyal ağlar üzerinden de kullanıcılara destek sağlamaya başlamışlardır (Geierhos, 2011). Kullanıcıdan gelen sorun, istek ve öneriler, destek personelinin etkin ve hızlı şekilde kullanabilmesi için, yardım bileti şeklinde tutulurlar. Tipik bir yardım bileti kullanıcıyı ve kullanıcının karşılaştığı durumu tanımlayan bilgileri ve/veya verileri barındırır.

Destek sürecinde, oluşturulan yardım biletleri değerlendirilip, kullanıcıya yanıt dönülmek üzere konu hakkında bilgi ve beceriye sahip teknik personele aktarılır. Özellikle büyük kuruluşlarda, kaynak tüketiminin verimli şekilde gerçekleştirilebilmesi için, kullanıcı destek sistemleri, gelen biletin gerektirmiş olduğu bilgi ve beceri düzeylerine göre seviyeler halinde düzenlenmiştir. ITIL (Information Technology Infrastructure Library)' in getirmiş olduğu standartlar içerisinde yer alan bu destek düzeyleri, 5 seviyeden oluşmaktadır. 0. Seviye kullanıcının destek personelinden yardım almadan, edindiği talimat ve yönlendirmeler ışığında çözüme ulaşmasıdır. Sık sorulan sorular ve tanıtıcı broşürler (izahnameler) bu seviyeye dâhildir. 1. Seviye basit bilgi ve beceri seviyesi gerektiren yardım biletlerinin çözüme ulaştırıldığı destek düzeyidir. Bu seviyede çözümlenemeyen yardım biletleri sırasıyla 2. ve 3. Seviye destek düzeylerine iletilirler. İlk üç seviyede çözülemeyen yardım biletleri 4. Seviye'ye ait olup, bu düzeye ulaşan biletler mevcut organizasyon haricindeki 3. Parti organizasyonlara çözülmeleri için aktarılırlar. Destek personel maliyeti ve personel bilgi seviyesi 1. Seviye'den 4. Seviye'ye yükseldikçe artar.(Wei ve diğerleri, 2007)

Destek sürecinin verimli şekilde gerçekleştirilebilmesi için 1.Seviye önemli rol oynar. Çünkü gelen yardım biletleri, konu ve içeriğine göre değerlendirilip, uygun uzman personel veya birime bu düzeyde aktarılır. Değerli destek kaynaklarının verimli şekilde taksim edilmesi ve müşteri memnuniyetini etkileyen yardım biletinin hızlı

şekilde yanıtlanması, biletin doğru birime aktarılmasına bağlıdır. Mevcut destek sistemleri gelen yardım biletlerini doğru birime yönlendirebilmek için yaygın olarak iki farklı yöntem etrafında gruplandırılırlar. İlki kullanıcıların uygun birim veya kategoriyi daha önceden belirlenmiş seçenekler içerinden işaretlemesine dayanır. Bu yöntemde kullanıcılar, kategoriyi yanlış şekilde işaretleyebilmektedirler. Bu sebeple biletler yanlış uzman personel veya birime yönlendirilebildikleri için, yanıt süresi artmakta ve destek kaynakları verimsiz şekilde kullanılmaktadır. Bir diğeri, yardım biletlerinin operatör(ler) tarafından değerlendirilerek uygun uzman personel veya birime atanmasına dayanır. Bu yöntem yorucu ve zaman alıcı olan pahalı insan çabası gerektirir. İnsan hatasından kaynaklanan yanlış yönlendirmeler olabilmektedir. Ayrıca el ile yapılan atama işlemi, ortalama yanıt süresini arttırmakta, bu durum müşteri memnuniyetini olumsuz etkilemektedir.

1.1 Tezin Amacı

Bu çalışmada, mevcut kullanıcı destek sistemlerini iyileştirmek için, makine öğrenmesi (yapay zeka) algoritmaları ve doğal dil işleme yöntemleri kullanılarak büyük organizasyonlara ait destek sistemlerinin ihtiyaçlarına cevap verebilecek yeni bir destek sistem mimarisi önerilmiş ve gerçekleştirilmiştir. Önerilen sistemde, yardım biletlerinin ilgili birime veya personele otomatik olarak atanması amaçlanmış, bunun yanı sıra tekrar eden yardım biletlerinin sistem tarafından otomatik cevaplanması hedeflenmiştir. Önerilen ve gerçekleştirilen bu sistem sayesinde, mevcut sistemlerde yardım biletlerinin ilgili destek personeline atanması için kullanılan hataya meyilli insan çabası gerektiren elle atama işlemi asgari düzeye indirgenmiş, yanıt süresi kısaltılarak müşteri memnuniyeti yükseltilmiş ve doğru atama işlemi yapılarak, sağlanan destek kalitesi arttırılmıştır. Ayrıca önerilen sistem değerli destek kaynaklarının verimli ve etkin şekilde kullanılmasına olanak sağlamıştır.

Bu çalışma yedi kısımdan oluşmaktadır. Bu kısımdan sonraki devam edecek kısımlar sırasıyla şu şekildedir: 2. kısımda daha önce yapılmış benzer çalışmalar incelenmiştir. 3. kısım tezin ana temasını oluşturan doğal dil işleme üzerine genel bilgiler içermektedir. 4. kısımda; sınıflandırma ve yapay öğrenme konularına değinilmiştir. 5. kısımda; çalışmanın uygulama aşamaları safhalar halinde incelenmiştir. 6. kısımda uygulamanın gerçekleştirilme aşamaları anlatılmıştır. 7. ve son kısımda değerlendirme ve öneriler yer almıştır.

2. ÖNCEKİ ÇALIŞMALAR

2.1 Destek Sistemleri İle İlgili Çalışmalar

Dar anlamda bakıldığında, önerilen çalışma kullanıcı destek sistemlerini iyileştirme ve geliştirme işlemidir. Akademik anlamda bu konuda yapılmış sınırlı sayıda çalışma mevcuttur. Bu çalışmalardan bazıları aşağı verilmiştir.

Düzensiz formda, doğal dil kullanılarak oluşturulmuş IT sorun biletlerinin, destek personelinin problem kaynağını belirleme ve probleme çözüm üretme hızını arttırmak amacıyla, otomatik olarak yeniden yapılandırılmasını sağlayan bir sistem önerilmiştir (Wei ve diğerleri, 2007). IT bilet metni, satır satır parçalara ayrılmış, her bir parçanın ne tür bilgi içerdiği CRF (conditional random field) makine öğrenmesi metodu kullanılarak tahmin edilmeye çalışılmıştır. Nihayetinde, biletin problem tipi ve problemi tanımlayan satırları düzenlenmiş halde destek personeline sunulmuştur. Önerilen yöntem %82 başarım göstermiştir.

Yardım biletlerini sınıflandırmak için, kitle kaynağı (crowd-sourcing) tarafından oluşturulmuş kural tabanlı bir yaklaşım önerilmiştir (Diao ve diğerleri, 2009). Basit sınıflandırma kuralların üretilebileceği ve yönetilebileceği sosyal ağ tabanlı xPad adında bir platform geliştirilmiştir. İnternet kullanıcılarından yardım biletlerini sınıflandırmak için basit kurallar üreterek katkıda bulunmaları istenmiştir. Çalışmada, kitle kaynağı tarafından üretilen kurallar ve gözetimli makine öğrenmesi sınıflandırma başarımları açısından karşılaştırılmış ve önerilen yaklaşımın işaretlenmiş verinin az olduğu durumlarda verimli sonuçlar verdiği görülmüştür.

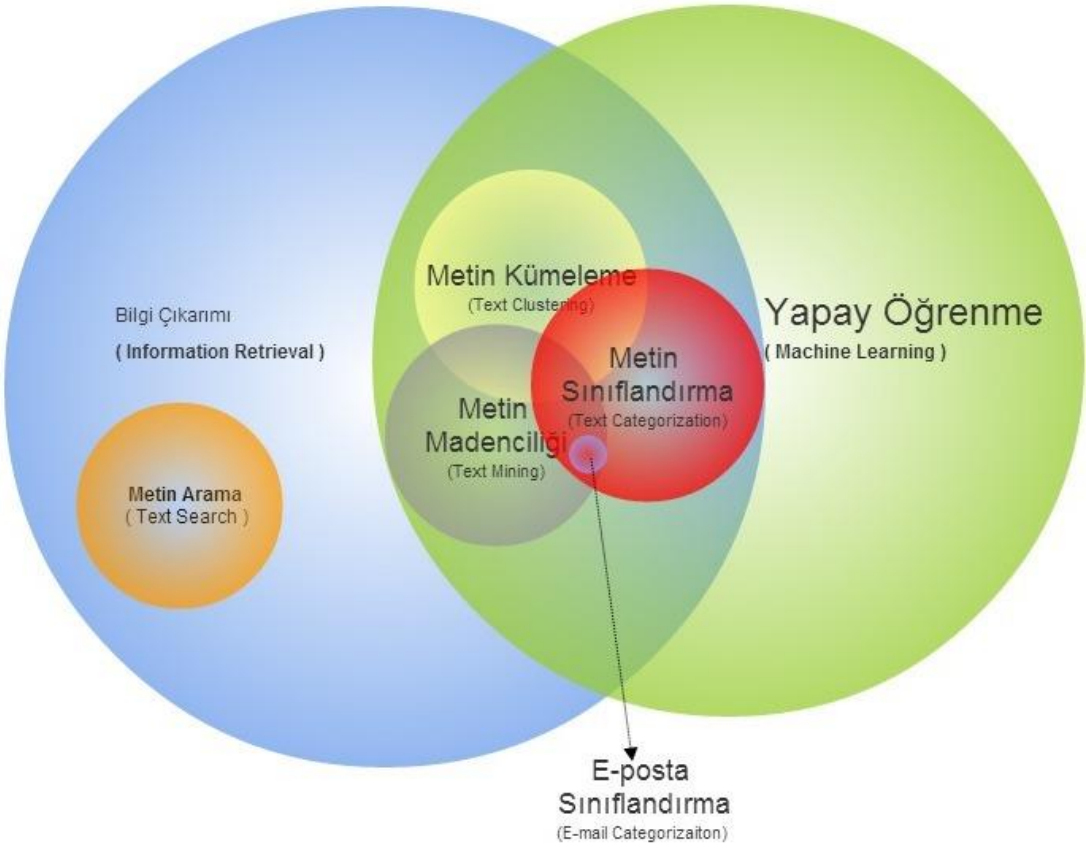
Bankacılık alanında IT yardım biletlerinin ilgili çözüm grubuna otomatik olarak atanması amacıyla metin madenciliği ve makine öğrenmesi algoritmaları kullanılarak bir sistem önerilmiştir (Sakolnakorn ve diğerleri). Otomatik atama işlemi sistem hata mesajları ve çözüm grubunu tanımlayan anahtar kelimeler arasındaki ilişki değerlendirilerek gerçekleştirilmiştir. Otomatik atama işlemi için WEKA aracı

kullanılarak karar ağaçlarının çeşitli algoritmalarının performansları karşılaştırılmış ve en iyi başarımı %93 ile ID3 algoritması göstermiştir.

Müşteriler tarafından oluşturulan yardım biletlerini içeren veri tabanı boyutunu, bütünlüğü koruyarak, azaltmak amacıyla çalışma yapılmıştır. (Weiss ve diğerleri, 2002). Bu yöntem sayesinde birbirine benzer yardım biletleri gruplandırılarak sorun çözüm aşamasında verimliliği arttırmak amaçlanmıştır. Genel itibariyle metin özetleme yöntemleri izlenmiştir. Önerilen yöntem, yaklaşık 1/3 oranında veri tabanı boyutu azaltma başarımı göstermiştir.

2.2 Doğal Dil İşleme Alanında Sınıflandırma Üzerine Yapılan Çalışmalar

Geniş anlamda bakıldığında, bu çalışmada yapılan, aslında bir metin sınıflandırma işlemidir. Metin sınıflandırma üzerine çok sayıda çalışma yapılmıştır. Bu konuda yapılan çalışmalar yapay zekânın bir alt kolu olan doğal dil işleme (NLP) alanında incelenmektedir. Şekil 2.1' de metin sınıflandırma ve diğer ilişkili alanlar küme halinde gösterilmiştir.



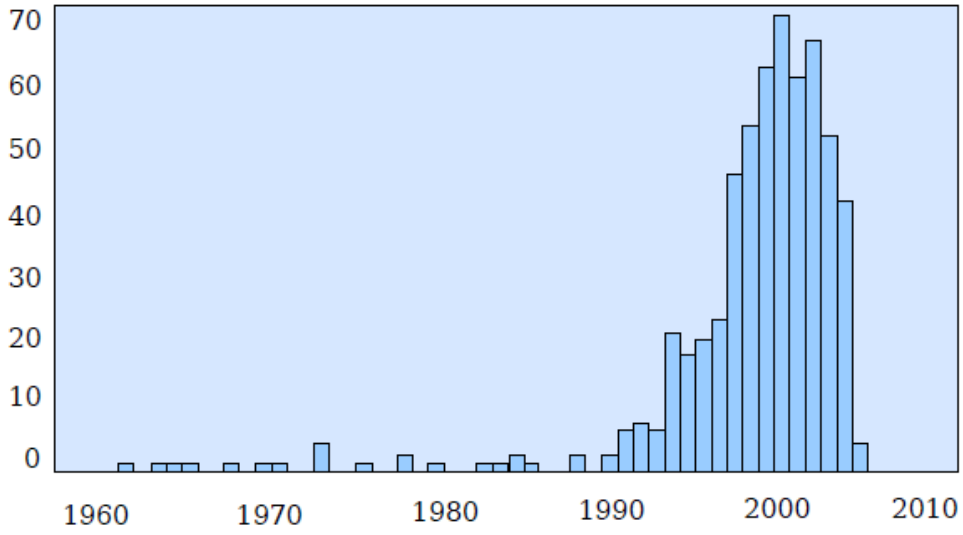
Şekil 2.1 : Metin sınıflandırma ve diğer ilişkili alanlar

Konuyla ilgili en eski referans; dilin analizini yapmak amacıyla hesaplama modellerini, istatistikler ile birleştirme fikrinden yola çıkılarak yapılan Luhn'un çalışmasıdır (Luhn, 1957). Bu çalışma, düşünce ve deneyimlerin dilde nasıl biçimlendiği konusunda felsefi yaklaşımlar kazandırmıştır. Bu yıllarda DDİ'nin alt konusu bilgi çıkarımı (information retrieval) alanında çalışan araştırmacılar, dokümanları ilgili gruplara ayırmak amacıyla, uzaklık öbekleme (clustering) üzerine çalışmışlardır. Bu dönemde işaretlenmiş verinin olmaması veya sınırlı miktarda olması araştırmacıları uzaklık ölçümleri (distance metrics) kullanarak nesneler arası benzerliği hesaplama üzerinde yoğunlaştırmıştır.

Yapay öğrenme modelleri kullanılarak yazın sınıflandırmak üzere yapılan ilk çalışma Maron' un 1961'deki çalışmasıdır. Maron bu çalışmasında Markov'un bağımsızlık varsayımını kullanarak olasılıksal sınıflandırma ile dokümanları indekslemeyi amaçlamıştır (Maron, 1961). Bu çalışmada özellik seçimi (feature selection) ve sıradan kelimelerin filtrelemesi (stop word filtering) gibi sınıflandırma tekniklerinde mihenk taşı konumunda olan önemli konular ele alındığı için kendinden sonra gelen pek çok çalışmaya referans olmuştur.

Daha sonra yapılan bir diğer çalışma Maron'u desteklemiş ve hesaplamalı modeller kullanarak dokümanları sınıflandırılabilmenin mümkün olduğu bir kez daha doğrulanmıştır (Borko ve Bernick, 1963). 70 ve 80'li yıllarda benzer çalışmalar yapılagelmiştir.

Şekil 2.2' de yazın sınıflandırma üzerine yapılan çalışmaların yıllık bazda dağılımını içeren grafik verilmiştir (Gabrilovich, 2007). Grafikte de görüldüğü gibi, 90'lı yıllarda yapay öğrenme çalışmalarına paralel olarak bu alanda yapılan çalışma sayısında artış gözlenmiştir. Bu artışta veri kümelerin (corpora) artması ve sınıflandırma performansının ölçülmesi için geliştirilen değerlendirme metrikleri etkili olmuştur. Bu alanda çalışmaların artması, dolaylı olarak araştırmacıları bazı yeni problemlerle karşılaştırmıştır. Yüksek boyutlu özellik vektörünün boyutunun düşürülmesi ve veri seyrekliği problemi bunlardan bazılarıdır. Bu problemler araştırmacıları yeni yöntemler aramaya itmiştir. Takip eden yıllarda, karar ağaçları (Fuhr & Buckley 1991, Quinlan 1993), kNN (Massand ve diğerleri 1992), Naïve Bayes sınıflandırıcısının çeşitli formları önerilmiştir (Maron 1961), (Duda & Hart 1973), (Fuhr, 1989).



Şekil 2.2 : Yıllar bazında yazın sınıfladırma üzerine yapılan çalışmaların dağılımı

Yapay öğrenme tekniklerinin geliştirilmesiyle birlikte sınıflandırma üzerine pek çok farklı çalışma gerçekleştirilmiştir (Apte, Damerau ve Weiss 1994, Fuhr ve Pfeifer 1994, Cohen 1995, Damashek 1995, Rilo 1996, Hull, Pedersen ve Schutze 1996, Joachims 1997). 1995 yılında Vapnik çekirdek ve kenar tabanlı sınıflandırma metodunu önermiştir (Cortes & Vapnik, 1995). 90'ların sonuna gelindiğinde çekirdek makineleri veya daha popüler adıyla SVM'ler yapay öğrenme üzerine çalışan araştırmacıları hayli etkilemiştir. Bu konuda yapılan çalışmalar SVM'lerin yazın sınıflandırma alanındaki başarısını doğrulamıştır (Joachims, 1998) (Yang ve Liu, 1999). Takip eden yıllarda SVM açık kaynak kütüphane paketleri (SVM^{light}[1], LibSVM[2]) olarak araştırmacıların kullanımına sunulmuştur.

Daha sonraki yıllarda sınıflandırma metotları kapsamlı bir şekilde karşılaştırılmış ve çeşitli yönlerden değerlendirilmiştir (Yang ve Liu, 1999) (Kotsiantis, 2007). Çizelge 2.1'de bu çalışmada elde edilen sonuçlar gösterilmektedir (Kotsiantis, 2007). Yüksek performanslı sınıflandırma modelleri ve yapay öğrenme araçlarının (SVM kütüphaneleri, MALLET[3] ve WEKA[4] gibi) uygulamalarda doğrudan kullanılabilecek

[1] Detaylı bilgiye http://svmlight.joachims.org/ adresinden ulaşabilirsiniz.
[2] Detaylı bilgiye http://www.csie.ntu.edu.tw/~cjlin/libsvm// adresinden ulaşabilirsiniz.
[3] Detaylı bilgiye http://mallet.cs.umass.edu/ adresinden ulaşabilirsiniz.
[4] Detaylı bilgiye http://www.cs.waikato.ac.nz/ml/weka/ adresinden ulaşabilirsiniz.

Çizelge 2.1 : Öğrenme algoritmalarının kıyaslanması

Parametre	Karar Ağaçları	Yapay Sinir Ağları	NaiveBayes	kNN	SVM	Kural Tabanlı Öğrenme
Genel Doğruluk	**	***	*	**	****	**
Öğrenme Hızı	***	*	****	****	*	**
Sınıflandırma Hızı	****	****	****	*	****	****
Eksik Özellik Değerlerine Yönelik Tolerans	***	*	****	*	**	**
İlgisiz Özellik Değerlerine Yönelik Tolerans	***	*	**	**	****	**
Gereksiz Özellik Değerlerine Yönelik Tolerans	**	**	*	**	***	**
Birbirine Bağlı Özelliklere Yönelik Tolerans	**	***	*	*	***	**
Gürültüye Yönelik Tolerans	**	**	***	*	**	*
Overfitting Problemi Karşısındaki Direnci	**	*	***	***	**	**

Parametre Alması Bakımından	***	*	****	***	*	***
Açıklanabilirliği	****	*	****	**	*	****
Artışlı Öğrenme (Incremental Learning) Girişimi	**	**	****	****	**	*
Niteliklerin Ayrık / İkili / Sürekli (Discrete/Binary/Continuous) Olma Durumuna Göre	****	*** (Ayrık Olmamalı)	*** (Sürekli Olmamalı)	*** (Doğrudan Ayrık Olmamalı)	** (Ayrık Olmamalı)	*** (Doğrudan Sürekli Olmamalı)

NOT: Bu tabloda, * en kötü performans ve **** en iyi performans olmak üzere öğrenme algoritmaları, parametrelere göre derecelendirilmiştir.

hale getirilmesiyle, sınıflandırma teknikleri temel alınarak geliştirilmiş çalışma sayısında artış gözlenmiştir. Literatürde sık karşılaşılan bazı çalışma konuları şöyledir; üslup analizi, duygu analizi, içerik tabanlı istenmeyen mail filtreleme, yazar tanıma, anahtar kelime ve bilgi çıkarımı, sözcükteki anlam karmaşasını giderme, varlık adı tanıma vb.

Bir sonraki başlık altında, metin sınıflandırma çalışmaları içerisinde çalışmamıza yakın olan e-posta sınıflandırma alanında yapılan çalışmalar incelenmiştir.

2.3 E-Posta Sınıflandırma Üzerine Yapılan Çalışmalar

İnternet erişilebilirliğinin artmasıyla birlikte e-posta popüler bir iletişim aracı haline gelmiştir. Elektronik ortamda bilgi paylaşımının hızlanmasına paralel olarak, e-posta trafiğinde fark edilir düzeyde artış gözlenmiş, artan bu trafik ile birlikte istenmeyen e-posta (spam) tespiti, e-postaların konularına göre sınıflandırılması ve dosyalanması, önemli mesajların tespiti ve öncelik verilmesi gibi işlemler önem kazanmıştır. Bu nedenle bu alanda çok sayıda çalışma yapılmış ve uygulama geliştirilmiştir. Bunlardan en eski ve yaygın olanı kural tabanlı çalışan mail araçlarıdır. Microsoft Outlook bu araçlara örmek olarak gösterilebilir. Bu tür araçlar kullanıcı odaklı olduğu için kullanıcıdan kural ve kelime listesi tanımlanması istenmektedir. Bu tip işlemlerin kullanıcıya bırakılması genellikle zaman ve kaynak kaybına neden olmaktadır.

Literatürde bu konuda yapılan çalışmalar genel olarak istenmeyen mail (spam) tespiti başlığı altında yoğunluk göstermektedir. İstenmeyen mail tespiti Yapay Zekâ (AI) ve Makine Öğrenmesi (ML) alanlarının özel ilgi alanı haline gelmiş ve bu konuda pek çok çalışma yapılmıştır. İstenmeyen mesajların tespiti prensip olarak yanlış pozitif ve yanlış negatif değerlerinin düşürülmesi esasına dayanan bir sınıflandırma algoritmasının kurulmasıyla gerçekleşir. Bu konuda çeşitli sınıflandırma algoritmaları önerilmiştir. Karar ağaçları (J48), Naïve Bayes sınıflandırıcılar, yapay sinir ağları (NN), en yakın komşu (nearest neighbors) algoritması, regresyon analizi ve destek yöney makineleri (support vector machine) bunlardan öne çıkanlardır. Bu sınıflandırma teknikleri çeşitli performans ölçümleri ve veri kümeleri üzerinde değerlendirilmiştir. Yapılan çalışmalarda istenmeyen mesajların filtreleme

metotlarının karşılaştırması gerçekleştirilmiş ve frekans tabanlı vektör modellerini önerilmiştir (Youn & McLeod, 2007) (Levis & Ringuatte, 1994).

Türkçe'de e-posta sınıflandırma üzerine sınırlı sayıda çalışma mevcuttur. Bunlardan biri çekimli dillere yönelik istenmeyen mail filtreleme çalışmasıdır (Ozgur ve diğerleri, 2004). Bu çalışmada yapay sinir ağları ve Bayes sınıflandırıcı üzerine çalışmışlardır. Özellik vektörü çıkarımında morfolojik analiz ve kök bulmaya vurgu yapmışlardır. Bir diğeri, spam mail filtrelemede n-gram model kullanarak verimliliği artırma üzerine gerçekleştirilen çalışmadır (Çıltık & Güngör, 2008). Bunların yanında, Türkçe'de yazın sınıflandırma üzerine yapılan bazı çalışmalar şu şekildedir; Türkçe gibi sondan eklemi dillerde yeniden düzenlenmiş istatistiksel dil modelleriyle sınıflandırma çalışması (Tantuğ, 2010), n-gram modellerin yazın sınıflandırma metotlarına etkisini görmek üzere yapılan çalışma (Güran ve diğerleri, 2009), yazından yazar, üslup ve cinsiyet tahminine yönelik yapılan çalışma (Amasyalı & Diri, 2006), haber sınıflandırması üzerine yapılan çalışma (Amasyalı & Yıldırım, 2004), kelime kökleri kullanarak yapılan yazın sınıflandırma çalışması (Çataltepe ve diğerlerinin, 2007), haber sınıflandırması için ön veri işlemlerinin incelendiği çalışma (Torunoğlu ve diğerleri, 2011).

3. DOĞAL DİL İŞLEMEYE GENEL BİR BAKIŞ

3.1 Doğal Dil İşleme Geçmişine Genel Bir Bakış

İlk bilgisayarların üretildiği tarihten bu yana düşünen, karar verebilen, kullanıcıların girdilerini yorumlayabilen akıllı cihazlar ve yazılımlar geliştirme fikri her zaman olagelmiştir. Ancak Alan Turing'in 1950'lerde bu konu üzerine ortaya attığı makinelerin insanları taklit edebileceği fikri, bilim insanlarını bu konuda motive etmiş ve pek çok ciddi çalışmayı beraberinde getirmiştir.

İkinci dünya savaşı sırasında büyük bir atağa geçen bilgisayar alanındaki çalışmalar, savaştan sonra iki temel çekim gücü etrafında toplanmış, otomatlar ve bilgi kuramı da dediğimiz bilginin nicelleştirilmesi etrafında yoğunlaşmıştır. 1948'de Shanon, Markov modellerini dilin otomatını oluşturmak için kullanmış, gürültülü kanal (noisy channel) ve decoding benzetmelerini yapmıştır. Ayrıca termodinamikteki entropi kavramını, DDİ' ye kazandırmıştır. 1957'de Noum Chomsky'nin tüm dillere uygulanabilir evrensel bir gramer yapısını önermesiyle dilin makinelere öğretilebileceği fikri daha da kuvvetlenmiştir. Chomsky'nin bu çalışmasından yola çıkan dil bilimci ve bilgisayarcılar ayrıştırma (parsing) algoritmaları geliştirmişlerdir. Geliştirilen bu algoritmalar ilk olarak, yukarıdan aşağıya (top-bottom) ve aşağıdan yukarıya (bottom-top) olarak tasarlanırken; daha sonrasında, dinamik programlama yöntemlerini de içerisinde barındırmıştır. İlk parsing sistemi Zeling Harris'in Dönüştürme ve Söylem Analizi Projesi (Transformation and Discourse Analysis Project) adı altında 1958-1959 yılları arasında geliştirilmiştir.

John McCarty, Marvin Minisky, Cloude Shanon ve Nothoniel Rochestern'den oluşan bir araştırma grubu 'Yapay Zeka' (AI, Artificial Intelligence) tabirini ilk defa kullanmışlardır. Kısa bir süre sonra ilk doğal dili anlayan sistemler geliştirilmiştir. Bu sistemler daha çok soru-cevap ve anahtar kelime araştırma üzerine olmuştur. 50'li yılların sonlarına doğru Bayes metodu optik karakter tanıma (OCR) üzerinde uygulanmıştır. 1966'da ALPAC (Automatic Languages Processing Advisor Committee) 'ın son 10 yılda yapılan çalışmaların beklenen neticeyi karşılamadığını

belirten raporundan sonra bu alanda yapılan çalışmalar durma noktasına gelmişse de, 1971' de Winograd'ın SHRDLU sistemi aslında o güne kadar kat edilen mesafeyi araştırmacılara göstermiştir. Bu sistem o güne kadarki görülmemiş zorlukta ve karmaşıklıktaydı. Blokları doğal dil ile verilen basit komutlar yardımıyla sıralayan bir robottu.*"Move the red block on top of the smaller green one"* yönergesini anlayıp icra edebilecek düzeydeydi. Winograd'ın bu çalışması cümleden anlam çıkarmak için ayrıştırma (parsing) problemiyle işe başlamanın doğru bir karar olduğunu göstermiştir. Bu projeden sonra dili anlama yeteneğine sahip bilgisayarlar üretmek için yapılan çalışmalar hız kazanmıştır. (Bu dönem için, Schank ve Albenson 1977 kavramsal bilgi üzerine yapılan çalışmalarına bakılabilir.)

1980'lerin sonunda Makine Öğrenmesi (Machine Learning) algoritmalarının DDİ' de uygulanmasıyla pek çok kayda değer çalışma gerçekleştirilmiştir. Ayrıca IBM Watson Araştırma Merkezi'nin de etkisiyle dil işlemede olasılıksal modellemenin kullanımında gözle görünür bir artış gözlenmiştir.

Bilgisayar alanındaki gelişmelerin etkisiyle 90'larda DDİ hızlı bir yükseliş yakaladı. Olasılıksal veri modelleri DDİ içerisinde standart duruma geldi. Bilgisayarların hızlarının, ram ve disk kapasitelerinin artması ile konuşma ve dil işleme alanında pek çok ticari girişim gerçekleşti. Buna dolaylı olarak DDİ (NLP) alanı farklı alt alanlara ayrılmıştır.

2000'li yıllarla birlikte araştırmacıların elinde DDİ' de karşılaştıkları problemler ve kullanılabilecek özelleşmiş veri kümeleri oluşmuştur. İstatistiksel makine öğrenmesinin DDİ'ye uyarlanmasında artış gözlenmiştir. SVM (Support Vector Machine), maksimum entropi ve Bayes modeli dil işlemede standart hale gelmiştir. Yüksek hızlı bilgisayarlar eğitim (training) ve yayılma (deployment) sürelerini beş yıl öncesiyle kıyaslanmayacak düzeyde düşürmüştür. Bu durum uygulama sayısını arttırmış, çeşitli veri kümelerinin oluşturulmasına sebep olmuştur. 2010'lu yıllara gelindiğinde bu veri kümelerinin denetimsiz (unsupervised) yöntemlerle oluşturulması artış göstermiştir.

3.2 Doğal Dil İşlemeye Genel Bir Bakış

Dil insanlar arasında iletişimi sağlayan kurallı sesler dizisidir. Dilin en küçük birimi olan seslerin birlikteliğiyle (veya tek heceli de olabilir) kelimeler, kelimelerin bir araya gelmesiyle tamlamalar (phrase), birden fazla tamlama veya kelimelerin bir araya gelmesiyle cümleler meydana gelir. Arkadaşımızla konuşurken, kitap okurken, sinemada film izlerken dil aracılığıyla etkileşimde bulunuruz. Bir dil bilimci olan Muharrem Ergin dili, "Dil, insanlar arasında anlaşmayı sağlayan tabii bir vasıta, kendisine mahsus kanunları olan ve ancak bu kanunlar çerçevesinde gelişen canlı bir varlık, temeli bilinmeyen zamanlarda atılmış bir gizli antlaşmalar sistemi, seslerden örülmüş içtimaî bir müessesedir." şeklinde tanımlamaktadır. Bu tanımda geçen dil tam olarak doğal dile karşılık düşmektedir. Çünkü doğal dilin yanı sıra özel amaçlara yönelik oluşturulmuş, grameri zaman içerisinde insanların günlük kabul ya da yönelimleriyle şekillenmemiş, tamamıyla insan eliyle yapılandırılmış diller de mevcuttur. C, Perl, Python vs programlama dillerine; Baleybelen[5], Elfçe[6], Esperanto[7] belli kişi veya kişiler tarafından oluşturulmuş yapay dillere örnek verilebilir. Araştırmacılar, insanlar arasında etkin bir şekilde kullanılan doğal dili, bilgisayarların kavrayıp üretebileceği bir hale getirmek için doğal dil işleme (NLP veya computational linguistic) adı altında çalışmalar yapmaktadır. Doğal dil işleme sabit algoritmalar içermeyip belirsizliklere sahip olduğundan NP (non deterministic polynomial time) problem olarak ele alınmaktadır.

Doğal dil işleme beş farklı bilim dalının kesiştiği bir noktada yer alır (Nabiyev, 2010) . Bunlar;

- Dilin doğasını anlamaya çalışan dil bilimi
- İnsan beyninin dili kullanma ve kavrama yetisi üzerinde çalışan ruhbilimi (psikoloji)
- Düşüncenin dil ile ilişkisini kavramaya çalışan felsefe

[5]Muhyi-i Gülşeni (1528-1604) tarafından itikadî bilgileri avamdan sakınıp, sırlamak ve Osmanlı'da ortak kültür dili yaratmak üzere 1574'de kurgulanan dünyadaki ilk yapay dil.
[6] Yazar J.R.R Tolkien tarafından kurgulanmış fantastik elf kavminin konuştuğu dil.
[7]Leh göz doktoru Ludwik Lejzer Zamenhof tarafından 1887 yılında oluşturulan yapay dil.

- İnsan beynini modellemeye çalışan yapay zekâ araştırmaları
- Ve son olarak doğal dil işleme uygulamaları geliştirmek isteyen bilgisayar bilimleri

Doğal dil işleme çalışmaları 1992 Cristal tarafından yapılan tarife göre beş temel başlıktan oluşmaktadır (Nabiyev, 2010). Ancak son yıllarda bu başlıklara bir altıncısı eklenmiş ve aşağıdaki hali almıştır.

1. Ses Bilimi (Fonetik)
2. Biçimbilim (Morfoloji)
3. Sözdizimsel Analiz (Syntax)
4. Anlamsal Analiz (Semantic)
5. Edimsel Analiz (Pragmatic)
6. Söylem Analizi (Discourse)

Çalışmanın bu bölümünün bundan sonraki kısmında sırasıyla kısaca bu alanlar ele alınacaktır.

3.2.1 Ses Bilimi (Phonetics)

Dilde konuşulan kelimeler konuşmanın yapı taşı diyebileceğimiz seslerden meydana gelmiştir. Sesler farklı milletlerde farklı şekillerde ifade edilebilir. Bazen tek bir harfle ifade edilebilirken, bazen birden fazla harfin bir araya gelerek oluşturduğu heceyle ifade edilebilir. Sesin harflerle ifade edilmesine, dile ait fonetik denir.

Fonetik dildeki seslerin insan tarafından nasıl oluşturulduğu, her sesin akustik olarak nasıl farklılıklar gösterdiği ve nasıl bu farklılıkların dijital ortama aktarılıp kullanılabilir hale getirilebileceği üzerinde çalışmalar yapan bilim dalıdır. Fonetik, konuşma sesinin ağızda nasıl üretilmesini inceleyen söyleyiş fonetiği ve konuşulan sesin akustik analizini yapan akustik fonetiği olmak üzere iki temel alan altında incelenebilir.

Yazım dilindeki kelimelerin ses dalgaları ile ifade edilmesi yani konuşma analizi (text to speech) için üzerinde çalışılan veri kümesindeki her bir yazılı kelimenin söylenebilir fonetik karşılığı, telaffuzunun (pronunciation) olması gerekir. Yine aynı şekilde konuşmanın yazılı hale getirilmesi yani konuşma tanıma (speech recognition) için de her bir akustik kelimenin fark edilebilir (ayırt edilebilir) telaffuzuna ihtiyaç duyulur.

3.2.2 Biçimbilim Analiz (Morphology)

Doğal dil işlemenin en temel seviyelerinden biri olan biçimbilimsel analiz, kelimelerin anlam taşıyan en küçük birimlerine ayrılıp analiz edilmesidir. Bu birimlere DDİ' de morfem denir.

Biçimbilimsel analizde üç temel görev amaçlanmaktadır (Nabiyev, 2010);

- Kelimelerin kökünü ayrıştırmak: Kelimelerin isim, fiil, sıfat, zamir, edat gibi kelime türlerinden hangisine ait olduğu tespit edilir. Her bir kelime türünün cümleye katacağı anlam bellidir. Dolayısıyla kelimenin kökünün türünü belirlemek cümleye kattığı anlamı tespit etmek açısında önemli ipuçları verir.

- Kelimelerin eklerini ayrıştırmak: Bir dilde mevcut olan olası ekler dilbilimciler tarafından tanımlanmıştır. Dolayısıyla tespit edilen ekin kök ile oluşturacağı kelime değerlendirilerek kelimenin o dile ait olup olmadığı tespit edilebilir.

- Eklerin türünü ayrıştırmak: Ekin kelimeye veya cümleye kattığı anlamı belirleyebilmek için ekin türünün tespit edilmesi gerekmektedir.

3.2.2.1 Morfemlerin Türleri

Bir morfem tek bir heceden oluşabileceği gibi birden fazla ses veya heceden de oluşabilir. Morfemler kök (stem) ve ek (affix) olmak üzere iki temel sınıfta incelenirler. Kök (main morpheme) kelimenin ana anlamını barındırır. Ek (additional morpheme) ise kelimeye cümle içinde görev yüklemek ve yeni kelimeler türetmek amacıyla köke eklenir. Ekler şekil itibariyle dört sınıftan oluşur.

Ön Ek (Prefix) : Kökün başına gelen eklerdir. Türkçe' deki pekiştirme eki ve İngilizce' deki olumsuzluk eki örnek gösterilebilir.

sap-sarı, kıp-kırmızı, ter-temiz gibi

un-expected gibi

Son Ek (Suffix) : Kökün sonunda gelen eklerdir. Türkçe genel itibariyle sondan eklemeli (agglutinative) bir dildir. Bir köke İngilizce'de sınırlı sayıda[8]son ek gelebilmesine rağmen, Türkçe'de teorik olarak sonsuz sayıda son ek gelmesi mümkündür. 'Çağdaşlaştırabildiklerimizdenmişsinizcesine' kelimesi 1 kök ve 12 son ek olmak üzere 13 morfemden oluşmaktadır. Türkçe'de ortalama olarak bir kelime 3 morfemden oluşmaktadır.

Circumfix[9] : Kökün aynı anda hem başına hem de sonuna gelen eklerdir. Türkçe ve İngilizce'de örneği yoktur. Almanca'daki geçmiş zaman eki (başına 'ge', sonuna 't' getirerek) örnek gösterilebilir.

sagen (söylemek)

gesagt (söyledi) gibi

Infix[4]: Kökün ortasına getirilen eklerdir. Türkçe'de örneği yoktur. İngilizce'de bazı argo kelimelerde karşılaşılır. Bir Filipin dili olan Tagolog dilinde rastlanır.

hingi (borç almak)

humingi (borçlu) gibi

3.2.2.2 Morfemlerin biraraya geliş şekilleri

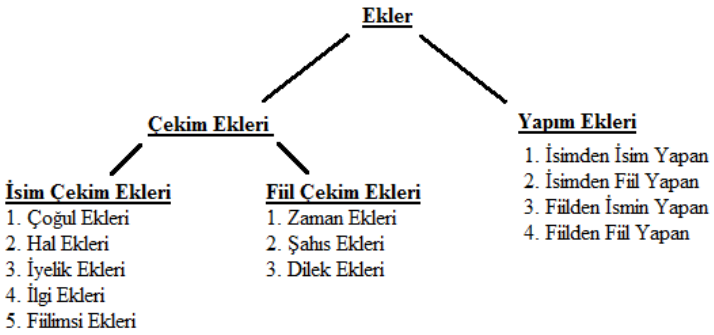
Morfemler birçok farklı yöntemlerle bir araya gelerek kelime oluştururlar. Doğal dil işlemede en çok karşılaşılan yöntemler şöyledir.

Çekimsel (Inflectional)

Kelimelerin cümle içerisinde farklı yerlerde ve görevlerde kullanılmasını sağlayan dil olayıdır. İngilizce'de isme eklenen çoğul (plural) ve sahiplik (possesive) ekleri, fiilde meydana gelen zaman (tense) değişiklikleri ve '-ing' takısı (gerund) örnek olarak gösterilebilir. Türkçe'de çekim ekleri bir kelime köküne veya gövdesine birden fazla kez eklenebilir. Soru eki 'mi' haricindeki tüm çekim ekleri kelimeye bitişik yazılır.

[8] Genellikle İngilizce'de bir kök 4'ten fazla son ek almaz.
[9]Türkçe'deraslamak mümkün olmadığı için terminolojide geçtiği şekilde yazılmıştır.

Türkçe'de ekler İngilizce'ye göre daha sistematik ve kurallıdır. Şekil 3.1' de Türkçe'ye ait yapım ve çekim ekleri verilmiştir.



Şekil 3.1: Türkçe'de çekim ve yapım ekleri

Yapımsal (Derivational)

Kelimelerden yeni kelime üretmek için kelimenin kök veya gövdesine ekin gelmesiyle oluşan dil olayıdır. Türkçe'de yapım ekleri çekim eklerinden önce gelir. İngilizce'de '-ness' ve '-er' ekleri örnek olarak gösterilebilir.

Bileşik şekilde (Compounding)

İki kelimenin bir araya gelip yeni kelime oluşturma olayıdır. Bir araya gelen kelimeler çoğu zaman kök halindedir. 'Çanakkale' ve 'doghouse' sırasıyla Türkçe ve İngilizce'de örnek olarak gösterilebilir.

Clitization

Türkçe'de görülmeyen ses olayıdır. Arapça ve İngilizce' de mevcuttur. Çizelge 3.1 İngilizce'deki bazı örneklerini göstermektedir (Jurafsky & Martin, 2009).

Çizelge 3.1 : Clitization örnekleri

Tam biçimi	Clitic	Tam biçimi	Clitic
am	'm	have	've
is	's	has	's
are	're	had	'd
will	'll	would	'd

Diğerleri

Tüm bu biçimbilimsel olaylarının yanında Arapça ve İbranice'de görünen şablon biçimbilimsel olayı aşağıdaki örneklerde gösterildiği gibidir.

ktb (yazmak) mektep (yazılan yer) Arapça

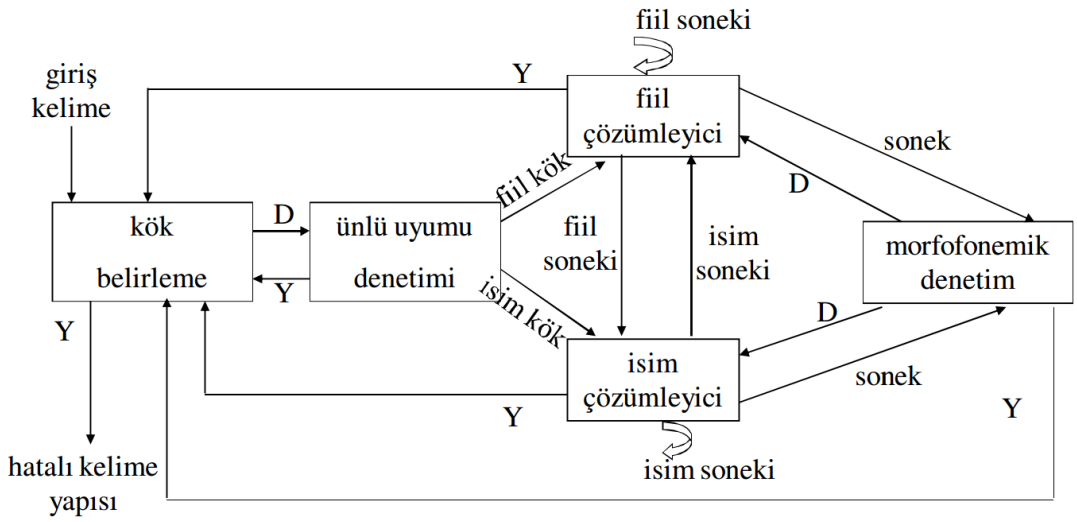
lmd (öğrenmek) limed (öğrendi) İbranice

Ayrıca bazı dillerde cinsiyet farklılığı vardır. Latin kökenli dillerden Fransızca, İspanyolca, İtalyanca iki cins barındırırken; Alman ve Slav dilleri 3 cins (Masculine, Femiline, Neuture) barındırmaktadır.

İngilizce'de özne yüklem uyumluluğu kesinlik göstermektir. Eğer özne tekil ise yüklem de tekil, eğer özne çoğul ise yüklem de çoğul olmak zorundadır. Türkçe'de bu kural İngilizce'ye nazaran daha esnektir.

3.2.2.3 Biçimbilimsel Analiz Yöntemleri

Biçimbilimsel analizin yapılmasında genellikle iki temel yöntem izlenmektedir. Bunlardan ilkinde, dilde mümkün olan tüm kelime biçimleri uygun morfoloji bilgileriyle beraber sözlükte tutulmaktadır. Bu durumda biçimbilimsel analiz uygun biçimin sözlükten bulunması ve uygun morfolojik sonuçların doğrudan alınmasıyla gerçekleşir. Bu yöntem sınırlı sayıda kelime biçimi içeren veri kümelerinde uygulanabilir. Ancak Türkçe gibi sonsuz sayıda kelime biçimi barındıran bir dilin modellenmesi için uygun değildir.



Şekil 3.2: Oflazer ve Bozşahin tarafından önerilen biçimbilimsel analiz

Diğeri, literatürde prosedürlü yöntem olarak anılmaktadır. Prosedürlü yöntemlerde sözlükte yalnız kelime kökleri saklanmaktadır. Kelimeyi kök ve eklerine ayırma işlemi

(parsing) yapılarak gerekli testler sonucunda biçimbilimsel sonuçlar elde edilmektedir. Şekil 3.2' de Türkçe için önerilen biçimbilimsel analiz akış şeması gösterilmiştir (Nabiyev, 2010).

3.2.3 Sözdisimsel Analiz (Syntax)

Bir dilde harfler uygun şekilde bir araya gelerek heceleri, heceler bir araya gelerek kelimeleri, kelimler anlamlı gruplar oluşturarak cümleleri meydana getirir. Kelimeler yığınının anlamlı bir cümle oluşturması için dilin gramerinin belirlediği kurallara göre bir birliktelik sağlaması gerekmektedir. Aşağıdaki örnek bunu güzel bir şekilde açıklamaktadır.

Beautiful to possible is more world build it a

Burada rastgele dizilişle bir araya gelmiş kelimelerin her birinin anlamını bilmemize rağmen, bütün olarak bir anlam ifade etmemektedir. Oysa aynı kelimeler aşağıdaki gibi sıralı bir şekilde bir araya geldiklerinde bir anlam ifade etmektedirler.

It is possible to build a more beautiful world

Türkçe İngilizce'ye göre daha serbest dizilimli bir dil olduğu için söz dizilimini açıklamak açışından burada İngilizce bir cümle örnek verilmiştir. Kelimelerin anlamlı bir birliktelik sağlamaları İngilizce'de daha çok, diziliş ve sırayla ilişkiliyken, Türkçe'de kelimeler aldıkları çekim ekleriyle anlamlı bir birliktelik oluşturulur.

Sözdizimsel analiz ile cümleyi oluşturan kelimelerden anlamlı cümle birimleri, tamlamalar veya anlamlı cümle parçacıkları tespit edilir. Bu tespit cümlenin beyan ettiği anlamı çıkarma (anlamsal analiz) açısından önemli rol oynar. Sözdizimsel analiz cümlenin yapısal tanımını oluşturabilmek için biçimbilimsel analizin sonuçlarını kullanır.

Cümle içerinde tek bir birimmiş gibi görev üstlenen kelime ve kelime gruplarına cümlenin bileşenleri (constituent) denir.

Küçük güzel evin penceresi[Özne] açıktı[Yüklem]

Yukarıdaki örnekte cümle özne ve yüklem olmak üzere iki öğeden oluşmaktadır.

Cümle bileşenlerini bulmak için kullanılan en yaygın yöntem, CFG (Context - Free Grammer) yöntemidir. Phrase Structure Grammer veya Backus Noum Form (BNF) olarak da isimlendirilebilir. 1900'lü yılların başında WilhelmWuth tarafından ortaya atılan dilin bileşenlerine ayrılarak modellenmesi fikri, 1956 yılında Chomsky tarafından formülize edilmiştir.

Cümle > İsim Tamlaması + Fiil

İsim Tamlaması > Sıfat Tamlaması + İsim

Sıfat Tamlaması > Sıfat + İsim

Sıfat > Sıfat + Sıfat

Sıfat > küçük | güzel

İsim >ev | pencere

Fiil > açıktı

Yukarıda *"Küçük güzel evin penceresi açıktı"* cümlesine ait CFG modellemesi verilmiştir. Burada kelimeyle ilişkilendirilmiş olanlar bitiş (terminal) sembollerini, cümleyle ilişkilendirilmiş olanlar başlangıç (start) sembolünü betimlemektedir. En genel bileşenden başlayarak en özele doğru daha farklı ifadeyle kökten yapraklara doğru bir süzgeçleme, yukarıdan aşağıya (top-down) veya en özelden başlayarak genele doğru aşağıdan yukarıya (bottom-down), yapraktan köke doğru çözümleme yaklaşımlarıyla cümle analiz edilir. Bazen her iki yaklaşım tek bir yöntem içerisinde de kullanılabilir. Bunların yanı sıra CYK ve Early gibi dinamik programlama algoritmaları da sözdizimsel analiz için kullanılmaktadır.

3.2.4 Anlamsal Analiz (Semantic)

Anlamsal analiz; morfolojik analiz ve sözdizimsel analiz sonuçlarını, kelime veya söz öbeği anlamlarını (knowledge source) kullanarak cümleden anlam çıkarmaya yönelik çalışır. Bilgi kaynağı (knowledge source); kelimelerin anlamını, dil bilgisi kurallarıyla olan ilişkiye göre barındırılan anlamı, söylemlerin yapısına göre barındırılan anlamı, konu üzerinde genel bilgiyi ve konunun genel kabul görmüş durumu hakkındaki bilgiyi içerir. Semantik analiz aşağıdaki şekilde betimlenebilir (Jurafsky & Martin**,** 2009**).**

Cümle Yapısı (Sentence Sturucture) + Kelime Anlamı (Knowlodge Source) = Cümle Anlamı (Sentence Meaning)

Anlamsal analizin tanımını yapmak için literatürde verilen örnekle konuyu izah etmeye çalışalım. Semantik analiz; bir tarifi uygulamak için tarifi elde etmenin yanında onu gereken tecrübe ve bilgiyle anlamlandırma işlemidir. Burada bahsi geçen tarifi elde etme; morfolojik ve sözdizimsel analiz sonuçlarıyla elde edilir.

Semantik analiz için kelimeler arasındaki anlamsal ilişki bilinmek zorundadır. Bu amaçla sözcüklerin birbiriyle olan ilişkilerinin oluşturulduğu ve tutulduğu veri tabanları mevcuttur.[10]

DDİ' de anlamsal analiz iki temel başlık altında ele alınmaktadır.

Sözdizimsel odaklı anlamsal analiz; bir fiilin hangi durum ve nesnelere uygulanabileceğiyle, fiil ve nesneler arasındaki ilişkiyle ilgilenir.

İsmail elmayı yedi. yemek(İsmail,Elma)

Sözcük odaklı anlamsal analiz; sözcüklerin anlamsal olarak sınıflandırılmasıyla (sesteş, anlamdaş vb) ilgilenir.

Elma bir meyvedir. Armut da bir meyvedir. Dolayısıyla elma ve armut bir meyvedir. Elma yenir. Dolayısıyla armut da bir meyve olduğu için armut da yenir.

3.2.5 Faydasal Analiz (Pragmatic)

Bir yazı veya diyalogu anlamak için yalnızca her bir cümleyi anlamak yetersiz kalır. Zira anlamsal bütünlük açısından cümleler arasındaki bağlantıyı da takip edebilmek gerekmektedir. Bu konuda yapılan çalışmalar DDİ 'de faydasal analiz olarak adlandırılmaktadır.

Ahmet sahaftan eski bir kitap aldı. Baş sayfası hayli yıpranmıştı.

---

[10] Semantik veri tabanı oluşturmak amacıyla düzenlenmiş CS Oyun bunlardan biridir. İlgilenenler http://www.kemikoyun.yildiz.edu.tr/commonsense/ adresinden ulaşabilirler.

Yukarıdaki örnekte baş sayanın kitaba ait bir unsur olduğu bilinmelidir.

Ayşe'nin yarın sabah sınavı var. Erken kalkması gerekmekte.

Yukarıdaki örnekte erken kalkması gereken kişinin Ayşe olduğu cümlelerin anlamsal bütünlüğü göz önüne bulundurularak tesbit edilmelidir.

Sema bu sabah çok mutluydu.

Bu cümlede Sema'nın bir özel isim olduğu anlaşılmalıdır.

Kar dün geceden bu yana yoğun bir şekilde yağmakta. Tüm köy yolları kapalı.

Bu cümle dizisinde yolların kapalı olmasının sebebinin yoğun kar yağışı olduğu çıkartılabilmelidir.

Sultan ISE101' den yüksek aldı.

Yukarıdaki cümlede Sultan'ın bir öğrenci ISE101'in bir ders ve yüksek aldı tabirinin ise yüksek bir not aldı şeklinde anlaşılması gerekmektedir.

3.2.6 Söylev Analizi (Discourse)

Birbiriyle ilişkili cümleler dizisine söylev denir. Bir diyalog içerisinde bir cümle genel itibariyle diğer cümlelerden izole edilerek anlaşılamaz. Cümlelerin birbirleri arasındaki ilişkiyi inceleyen DDİ aşamasına söylev analizi denir. Konuşmadaki ses tonu ve vurgu, benzetme atasözü ve deyim kullanımı, zamir kullanımı gibi konular söylev analizi altında incelenmektedir. Söylev analizi doğal dil işlemenin üst seviyelerinde yer alır.

4. SINIFLANDIRMA

Bu kısımda, konuya bütün olarak bakabilmek açısından sınıflandırma, yapay öğrenme ve bazı yapay öğrenme teknikleri bahsedilecektir.

4.1 Sınıflandırma

Sınıf, özelliklerine göre nesnelerin yerleştirildiği kategorilerin her biridir ve ilişkili, benzersiz tekil bir etiket ile adlandırılır. Örneğin etoburlar sınıfı et ile beslenen canlıları barındırmaktadır. Uygun nesnelerin bir araya getirilerek etiketlenmesi işlemine sınıflandırma denir.

X, uzay

Y, olası etiketler uzayı

x: $x \in X$, her hangi bir nesnenin çeşitli niteliklerine ait değerlerden oluşan özellik örnek uzayı

y: $y \in Y$, etiket kümesi içerisindeki etiket

olmak üzere;

f: $X \rightarrow Y$, sınıflandırma fonksiyonudur.

Nesneler çeşitli niteliklerine ait olası ölçümler kümesinden kendilerine ait değerler ile temsil edilmektedir. Somut bir örnek ile açıklamak gerekirse;

> *Alt ve üst çenelerinde ikişer tane sürekli büyüme eğiliminde olan kesici dişler bulunur. Köpekdişleri yoktur. Gözler başın her iki yanında toplandığından hem önlerini hem arkalarını aynı derecede görebilirler. Çok iyi koşar, tırmanır, sıçrar ve yüzerler.*[11]

Yukarıda memelilerin sınıfının alt sınıfı olan kemirgenler tanımlanmıştır. Tanımlama yapılırken bu türe ait özellikler nitelendirilmiştir. Sınıflandırma yapılırken nesnelerin ait

[11]Vikipedi'den alınmıştır

olduğu özellik değerleri göz önünde bulundurularak karar verilir. Farklı etiketler için farklı özellik kümeleri esas alınır. İdeal sınıflandırma işlemi azami fayda sağlayacak özellikler kümesi ile gerçekleştirilir. Bir nesne bir veya birden fazla ayrık olmayan sınıfa dâhil olabilir. Ayrık sınıf, olası ortak elemanları olmayan sınıfları ifade etmektedir. Örneğin, bir meyve hem elma, hem de portakal sınıfına dahil olamaz ancak, hem ekşi hem de sulu sınıfına dahil olabilir. Eğer uzayımızı iki ayrık sınıfa ayırmaya çalışıyorsak, ikili sınıflandırma söz konusudur. İstenmeyen mail filtreleme işlemi ikili sınıflandırmaya örnektir. Bu tür sınıflandırmalarda Aristo mantığı geçerlidir. Ya sınıfa dâhildir ya da değildir. Literatürde karşılaşılan sınıflandırma türleri Çizelge 4.1'de verildiği şekildedir. Buradaki P işareti çoklu etiketi betimlemektedir.

Çizelge 4.1 : Sınıflandırma tipleri

Uzay(X)	Etiketler(Y)	Fonksiyon(*f*)	Sınıflandırma Türü
meyve	{Ham,Olgun}	X→Y	tek etiketli, iki sınıflı (single-label,binary)
meyve	{Elma,Portakal,Şeftali}	X→Y	tek etiketli, çok sınıflı (single-label,multi-class)
meyve	{Ekşi,Sulu,Kabuklu}	X→PY	çok etiketli, çok sınıflı (multi-label,multi-class)
meyve	{Ekşi,Kabuklu}	X→PY	çok etiketli, iki sınıflı (multi-label,binary)

Bu çalışmada yapılan yardım biletlerinin sınıflandırılması işlemi tek etiketli, çok sınıflı (single-label, multi-class) sınıflandırma türüne girmektedir.

Sınıflar derecelendirme yapılarak veya sayısal değerler ile betimlenerek yumuşatılabilir. Bulanık mantık uygulamaları yumuşatılmış sınıflandırmaya örnektir.

4.2 Yapay Öğrenme

Bilgisayar ile bir sayının bölenlerini bulmak istersek, matematikteki mevcut olan bölen bulma yöntemini algoritmaya dönüştürmemiz yeterli olacaktır. Ancak gerçek hayatta karşılaştığımız pek çok problem için bilinen bir algoritma mevcut değildir. Örneğin;

ülkelerin gelecek beş yıl içerisindeki ekonomik durumlarını tahmin etmek, algoritması olmayan bir problemdir.

Olaylar, bir süreç içerisinde birbiriyle ilişkili ardışıl durumlardan meydana gelir. Bu süreci kesin bir şekilde tahmin edemesek de rastgele olmadığını biliriz. Bir sonraki durum hakkında sezgisel bir tahmin yürütebiliriz. Somut bir örnekle açıklamak gerekirse; giyim mağazaları insanların giyinme alışkanlıkların mevsimlerle ilişkili olduğunu varsayarak, reyonlarını mevsimlere göre düzenlemekte ve başarılı sonuçlara ulaşmaktadırlar.

Günümüz teknolojisinde çok büyük veri kümeleri saklanabilmekte, işlenebilmekte ve ağ üzerinden fiziksel olarak çok uzak mesafeler arasında paylaşılabilmektedir. Bu durum bizlere öğrenebilen sistemler tasarlama olanağı sağlamaktadır. Yapay öğrenme, bilgisayarların daha önceki veri ya da geçmiş deneyimi kullanarak başarımları arttıracak en iyi parametrelerle bir model oluşturulmak üzere programlanmasıdır (Alpaydın, 2011). Yapay zekânın alt dallarından biridir. Teorik bilgisayar bilimini ve istatistiksel metotları kullanır.

Yapay öğrenme, tahmin etme, sınıflandırma ve tanıma işlemleri gibi pek çok alanda yaygın şekilde kullanılır. Konuşma tanıma, hava durumu tahmini, yüz tanıma, yazın sınıflandırma, istenmeyen mail tespiti bu alanlardan sadece bir kaçıdır.

4.3 Yapay Öğrenme Teknikleriyle Sınıflandırma

Daha önce de bahsedildiği gibi yapay öğrenme, var olan veriler ile uygun model oluşturarak görülmemiş veriyi tespit etmeye çalışmaktadır.

$D = \{(x_1, y_1),(x_2, y_2), \ldots,(x_n, y_n) \mid (x_i, y_i) \in D : t(x_i) = y_i\}$

D kümesindeki her bir (x_i, y_i) çifti için, x_i örnek uzayına ait hedef fonksiyonu $(t(.))$ bir y_i değerine eşittir. Yapay öğrenmede bu tarz bir kümeye işaretlenmiş veya etiketli veri kümesi denir. İşaretlenmiş veri kümesi iki amaç için kullanılır;

1. Öğrenme modelini oluşturmak için eğitim kümesi olarak kullanılır.
2. Oluşan modeli sınamak için test kümesi olarak kullanılır.

Öğrenme teknikleri kullanılan eğitim kümesine göre kategorize edilmektedir. Eğer eğitim kümesi, işaretlenmiş veriden oluşuyor ise gözetimli öğrenme (supervised learning), işaretlenmemiş veriden oluşuyor ise gözetimsiz öğrenme (unsupervised learning) olarak tanımlanmaktadır. Literatürde gözetimsiz öğrenmeye öbekleme (clustering) de denilmektedir. Görüntü sıkıştırma algoritmaları öbeklemeye örnektir. Bunların yanı sıra öğrenme işleminde uygun sınıflandırıcıları oluşturmak için sınırlı sayıda işaretlenmiş veri kullanılıyorsa yarı gözetimli öğrenme (semi-supervised learning) şeklinde adlandırılır. Eğer yalnızca başlangıçta bir miktar işaretlenmiş veri ile başlayıp daha sonrasında çıkan sonuçlarla geri beslemeli bir şekilde devam ediyorsa bu öğrenme türüne de önyüklemeli öğrenme(boot strapping learning) denir.

Devam eden başlıklar altında, sınıflandırma için kullanılan öğrenme yaklaşımlarına kısaca değinilecektir.

4.3.1 Olasılıksal yaklaşım

Yapay öğrenme yaklaşımlarının en eskilerindendir. Bayes kuralına dayanan sınıflandırıcı metotları bu yaklaşıma dâhildir. Naive Bayes sınıflandırıcısı sık karşılaşılan olasılıksal sınıflandırıcı metodudur.

4.3.2 Olasılıksal olmayan yaklaşım

Literatürde olasılıksal olmayan sınıflandırma teknikleri de mevcuttur. Bu tekniklerden en basit ve etkili olanı çekirdek makineleri (kernel machine) olarak da anılan destekçi yöney makineleridir (Support Vector Machine, SVM). Karar ağaçları (decision trees), en yakın k komşu (k nearest neighbor) ve perceptron algoritmalarının çeşitli türleri bu yaklaşıma dâhil olan diğer sınıflandırma metotlarıdır.

4.3.3 Ardışıl yaklaşım

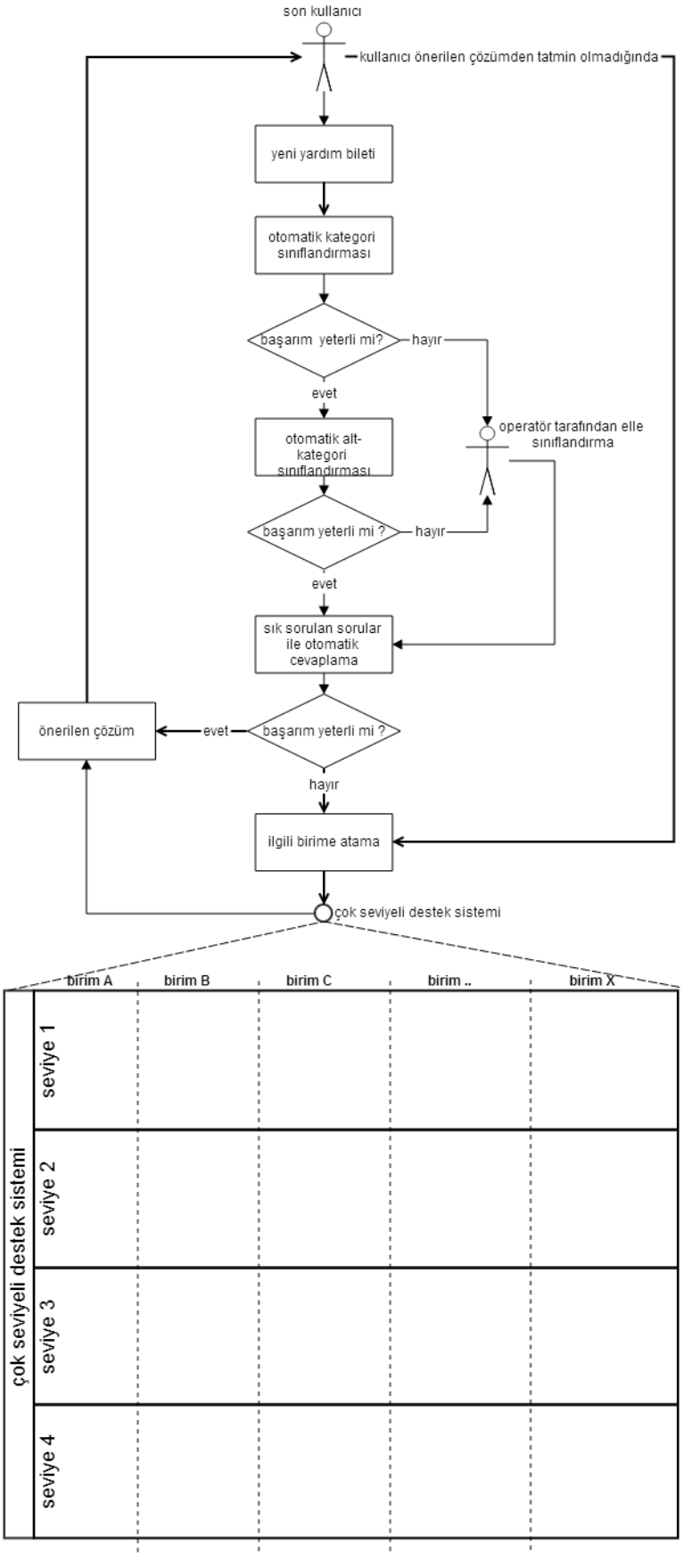
Birbiriyle ilişkili halde sıralı örneklerden oluşan veri kümesi üzerinde model oluşturan öğrenme yaklaşımlarıdır. DDİ' de kelime parçacıklarının gramer yapısını etiketleme (part of speech tagging) ve konuşma tanımada ses işaretleme (phone labelling) bu yaklaşıma dayanır. Saklı Markov Modelleri (Hidden Markov Models - HMMs) ve Şartlı rastgele Alanlar (Conditional Random Fields - CRFs) literatürde sık karşılaşılan ardışıl yaklaşımlı sınıflandırma yöntemleridir.

5. ÖNERİLEN YARDIM BİLETİ SİSTEMİ VE UYGULAMA YÖNTEMLERİ

5.1 Önerilen Sistem Mimarisi

Bu çalışmada, mevcut kullanıcı destek sistemlerini iyileştirerek müşteri memnuniyetinin arttırılması ve değerli destek kaynaklarının verimli ve etkin şekilde kullanılması amaçlanmıştır. Makine öğrenmesi (yapay zekâ) ve metin işleme algoritmaları kullanılarak büyük organizasyonlara ait destek sistemlerinin ihtiyaçlarına cevap verebilecek yeni destek sistem mimarisi önerilmiştir. Önerilen sistemde, son kullanıcıdan gelen yardım biletleri ilgili destek personeline veya birimine atanmak üzere peş peşe iki sınıflandırma işlemine tabi tutulmuştur. İlk sınıflandırma işlemi bilete ait ilgili kategoriyi tespit ederken, ikici sınıflandırma işlemi tespit edilen kategori altındaki alt kategoriyi belirlemektedir. Sınıflandırma başarımı daha önceden belirlenen eşik değerinin altında kaldığı takdirde, yardım bileti ilgili kategori ve/veya alt kategoriye sınıflandırılmak üzere operatöre gönderilir. Önerilen sistem, destek kaynaklarının azami derecede verimli şekilde kullanılmasını hedeflemektedir. Bu amaçla, yardım biletleri ilk olarak belirlenen alt kategoriye tanımlanmış sık sorulan sorular kullanılarak otomatik olarak cevaplanmaya çalışılır. Eğer buradan kullanıcıyı tatmin edici bir sonuç alınmaz ise yardım bileti belirlenen kategori ve alt kategoriye ait ilgili destek personeline veya birimine cevaplanması için atanır. Gelen yardım biletine otomatik olarak veya destek personeli tarafından önerilen cevap kullanıcıya iletilir. Eğer kullanıcı önerilen bu cevap ile tatmin olmaz ise yeniden değerlendirilmek üzere ilgili birime yeniden atama işlemi gerçekleştirilir. Şekil 5.1 önerilen sistem mimarisini göstermektedir.

Yardım biletlerinin ilgili kategori ve alt kategoriye atanması temelde tek etiketli, çok sınıflı bir metin sınıflandırma problemidir. Bu problemle ilgili literatürde çeşitli algoritmaların ve özellik çıkarım tekniklerinin kullanıldığı çok sayıda çalışma mevcuttur. Önerilen sistem her ne kadar dilden bağımsız olsa da, yardım biletleri Türkçe, İngilizce gibi doğal dil ile yazılmış metin içerdiği için sistemin uygulama aşamasında dile özel bazı işlemler yapmak gerekmektedir.

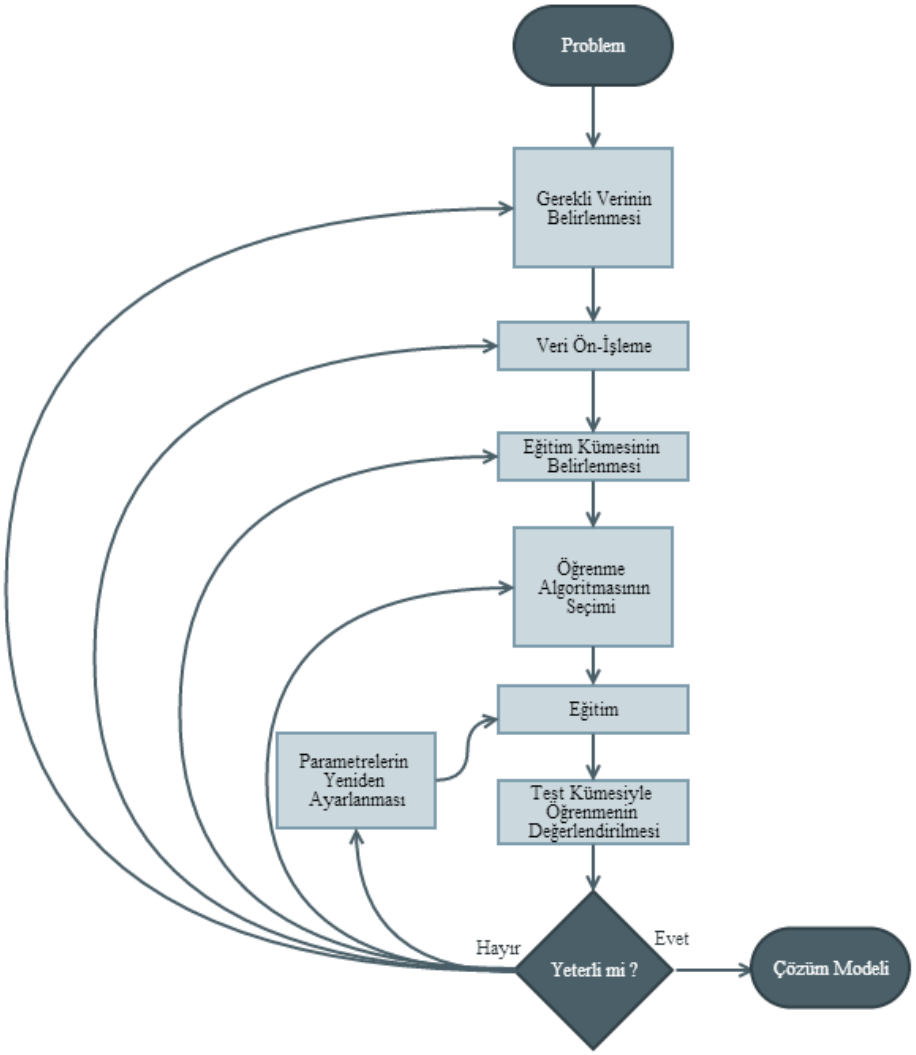


Şekil 5.1 : Önerilen sistem mimarisi

Önerilen mimaride yer alan sınıflandırma işlemleri için gözetimli yapay öğrenme teknikleri kullanarak öğrenme modelinin oluşturulması ve yardım biletlerinin daha önceden verilen cevaplar kullanılarak otomatik cevaplanması devam eden başlıklar altında detaylı şekilde ele alınmıştır.

5.2 Öğrenme Modelinin Oluşturulması

Gözetimli yapay öğrenme tekniklerinde amaç, daha önce etiketleri ve özellikleri verilmiş veriler arasındaki ilişkiyi inceleyerek bir öğrenme modeli geliştirmek ve oluşturulan bu model sayesinde ilk defa görülen etiketlenmemiş veriye olası etiketi atamaktır. Gerçek hayatta karşılaşılan bir problem karşısında geliştirilen gözetimli yapay öğrenme süreçleri şekil 5.2'de verilmiştir. (Kotsiantis, 2007)



Şekil 5.2 : Gözetimli yapay öğrenme tekniklerinde izlenen işlem süreçleri

5.2.1 Veri kümesi ve özellik vektörü

Öğrenme işleminin ilk aşaması veri kümesinin oluşturulmasıdır. Daha önce de belirtildiği gibi nesneler sahip oldukları özellikleriyle betimlenirler. Nesnelerin ait olduğu vasıflardan öğrenme için uygun bilgiyi taşıyan özellik ve/veya nitelikler (features, attributes) uzman tarafından belirlenir ve veri kümesi oluşturmak üzere toplanır. Eğer uygun nitelikler kestirilemiyor ise veriye ait ölçülebilen tüm alanlar toplanır. Buna literatürde kaba kuvvet veri toplama (brute force data collection) denir. Tüm niteliklerin toplanması yani kaba kuvvet yöntemi veri boyutunun gereksiz olarak artmasına sebep olmakla beraber, toplanan veri gürültü içerdiğinden dolayı algoritma performansını da olumsuz yönde etkilemektedir.

Nesnelerin bazı vasıfları, oluşturulmak istenen model için uygun bilgi içermeyebilir. Örneğin hastalık teşhisi için hastanın yaş durumu önemlidir. Kişinin doğum tarihi bilgisi doğrudan fayda sağlamayıp yaşı hesaplamak için kullanılır. Özellik uzayından (vector space) var olan vasıfların kullanılarak faydalı bilgi içeren yeni vasıflar elde edilmesi işlemine özellik inşası (feature construction) denir. Özellik inşası yaparken dönüştürülecek özellikler eşitlik ve eşitsizlik operatörleri (=,≠,<,>), aritmetik operatörler (+,-,×,÷), dizi operatörleri (maksimum, minimum, ortalama, vb) kullanılarak faydalı forma dönüştürülür. Bazı durumlarda özellik uzayı daha az karmaşıklığa sahip alt uzaya indirgenebilir. Bu işleme özellik seçimi (feature selection) denir. Örneğin, bir çalışmada uydu görüntüsünden çam ağacı yetişen bölgelerin tespiti amaçlanmaktadır. Bilindiği üzere iğne yapraklı bitkiler tüm mevsim yapraklı olurlar ve yaprakları yeşildir. Öyleyse tüm görüntüyü taramak yerine görüntünün histogramında yeşil tonun ağırlıklı olduğu bölgelerin taranması, karmaşık haldeki veriyi daha basit hale indirgemekte ve algoritma performansını olumlu yönde etkilemektedir.

5.2.2 Veri ön-işleme aşamaları

Nesnelerin ait olduğu vasıflarından oluşan veriyi sayısal hale dönüştürmek için veriler bir takım ön işleme (pre-processing) aşamalarından geçmektedir. Ön-işleme aşamaları problemin cinsine ve kullanılan veri kümesine göre değişiklik gösterebilmektedir. Görüntü işlemede görüntüyü oluşturan piksel değerleri özellik değerleriyle ilişkiliyken, metin işlemede metni oluşturan terimlerin terim sıklıkları özellik değerleriyle ilişkilidir.

Metin halindeki veriyi sayısal şekilde ifade edebilmek için bir takım işlemler gerçekleştirilir. Harris bunun için kelime torbası (bag of words) modelini önermiştir

(Harris, 1954) Bu modele göre, her bir doküman sözdizimi ve dilbilgisi kuralları göz önüne alınmaksızın düzensiz şekilde yığılmış kelime veya terimler halinde temsil edilebilir. Her bir terim bulunduğu doküman içerisinde kendine ait bir frekans değerine sahiptir ve sahip olduğu bu değer bir özellik (feature) olarak kabul edilmektedir. Bag of word yaklaşımı doküman sınıflandırma başta olmak üzere, görüntü işleme gibi pek çok alanda yaygın şekilde kullanılmaktadır.

Devam eden başlıklar altında, kelime torbası (bag of word) yaklaşımından yola çıkarak metnin sayısal özellik vektörüne dönüştürülebilmesi için gerekli aşamalardan ve karşılaşılan bazı problemlere getirilen çözüm yöntemlerinden bahsedilmiştir.

5.2.2.1 Terimlerin ağırlıklandırılması

İkili terim ağırlıklandırma (Boolean term weighting)

Özellik vektörü içerisindeki terimleri ağırlıklandırmanın en basit yolu, onları doküman içerisinde var olma durumuna göre 0 (sıfır) veya 1 (bir) ile ilişkilendirmektir. Eğer terim doküman içerisinde mevcut ise 1, değil ise 0 olarak ağırlıklandırılır. Bu yöntemde tüm terimler eşit önem derecesine sahiptir.

Terim sıklığı (Term frequency)

Bir doküman içerisinde, diğer terimlere göre daha sık geçen bir terimin önemi daha fazladır. Bu terim sıklığı (term frequency) ile ifade edilir. *tf* şeklinde gösterilir. Terim sıklığı bir terimin, dokümanda geçme sayısının o dokümanda en fazla geçen terimin sayısına oranıdır **(5.1)**.

$$tf(t_x, d_x) = \frac{f(t_x, d_x)}{max(f(t_0, d_x), f(t_1, d_x), f(t_2, d_x), \dots, f(t_n, d_x))} \qquad \textbf{(5.1)}$$

Örneğin *"okul"* kelimesi; A dokümanı içerisinde 10 defa geçmekte ve A dokümanı içinde en fazla geçen terimin doküman içerisinde geçme sayısı 45, B dokümanı içerisinde 3 defa geçmekte ve B dokümanı içende en fazla geçen terimin doküman içerisinde geçme sayısı 30 olsun. Bu durumda A ve B dokümanlarında, *"okul"* kelimesinin terim frekansı aşağıdaki gibidir.

$$okul_{A\ dokümanı\ için\ terim\ frekansı} = \frac{10}{45} = 0{,}222$$

$$okul_{B\ dokümanı\ için\ terim\ frekansı} = \frac{3}{30} = 0{,}1$$

Terim sıklığı - ters doküman sıklığı (Term frequency - inverse document frequency)

Dokümanlar kümesinde daha az sayıda dokümanda görülen bir terimin ayırt edici özelliği daha fazladır. Bu durum devrik doküman sıklığı (inverse document frequency) ile ifade edilir. Literatürde IDF şeklinde gösterilir. Doküman kümesindeki tüm dokümanların sayısının (n), terimin geçtiği doküman sayısına(n_x) oranının logaritması olarak hesaplanır **(5.2)**. Bu değer kıyaslama için kullanılacağı için logaritmanın tabanının önemi yoktur. Çoğunlukla *e*, 2 veya 10 olarak alınmaktadır. Logaritma kullanılmasının sebebi üssel fonksiyonun tersi yönde bir hesap yapmaktır.

$$idf(t_x) = \log(\frac{n}{n_x}) \qquad \textbf{(5.2)}$$

Yukarıdaki belirtilen iki sebep göz önünde bulundurulduğunda; bir dokümanda çok geçen, ancak diğer belgelerde daha az görülen bir terimin ağırlığı daha fazladır. Bir terimin tf-idf değeri terim sıklığı ve devrik doküman sıklığıyla doğru orantılıdır **(5.3)**.

$$tfidf(t_x, d_x) = tf(t_x, d_x) * idf(t_x) \qquad \textbf{(5.3)}$$

5.2.2.2 Veri seyrekliği problemi ve önerilen çözümler

Türkçe, Japonca, Fince gibi sondan eklemi dillerde bir kelime pek çok farklı formda bulunabilir. Çizelge 5.1'de *"dolap"* kelimesinin sık karşılaşılan bazı formları yer almaktadır.

Çizelge 5.1 : "dolap" kelimesinin sık karşılaşılan formları

Aldığı Ek	Bulunduğu Form
Yalın hali	dolap
Yönelme hali	dolaba
Yükleme hali	dolabı
İlgi eki almış hali	dolabın
1. Tekil sahiplik eki ile	dolabım
1. Çoğul sahiplik eki ile	dolabımız
Çoğul eki ile	dolaplar

Türkçe ‘de teorik olarak bir kök sonsuz sayıda ek alabilir. Yani metin içerisinde bir terimin pek çok farklı formdaki haliyle karşılaşmak mümkündür. Bu durum işlem yükü ve karmaşıklığını arttıracağı gibi, veri seyrekliği problemine de neden olacaktır. Ayrıca özellik vektör boyutu hayli artacaktır. Aşağıdaki alt başlıklarda duruma çözüm olarak önerilen yaklaşımlar ele alınmıştır.

Kök bulma (Stemming) yaklaşımı

Metin içinde geçen terimin kökü biçimbilimsel analiz ve takip eden morfolojik belirsizlik giderme işlemiyle tespit edilerek özellik vektörüne eklenir. Amaç veri seyrekliği problemini gidermek ve özellik vektörü boyutunu düşürmektir.

Biçimbilim Analiz (Morphology): Biçimbilimsel analiz DDİ’de kelimelerin kök, ek ve morfemlerinin tespitinde kullanılan temel adımlardan biridir[12].

Belirsizlik giderimi (Disambiguation): Biçimbilimsel analiz bir kelimeyi birden fazla şekilde etiketleyebilir. Aşağıda ‘araba’ kelimesinin olabilecek morfolojik çözümleri verilmiştir.

arap (arap +Noun+ A3sg+ Pnon+ Dat)
araba (araba+Noun+ A3sg+ Pnon + Nom)

Biçimbilimsel belirsizlik giderici kelimeye, metin veya cümle içerisindeki konumuna göre uygun morfolojik çözümün atanmasını sağlar.

İlk n karakter (First n character) yaklaşımı

Metin içerisindeki her kelimenin biçimbilimsel analizi çoğu zaman külfetli olmaktadır. Bu yüzden kök bulma yaklaşımına alternatif olarak terimin ilk n karakteri özellik vektörüne dâhil edilmektedir. Bu yaklaşım uygulama için kolaylık sağlarken doğruluktan ödün vermektedir. Türkçe'de nadiren rastlanan ön ek barındıran kelimeler ve anlamı olumsuzlaştıran "-sız" gibi ekli kelimeler için sorun teşkil edebilmektedir. “Sımsıcak”, “anti-sosyal”, “namert”, “merhametsiz” bu sorun teşkil eden kelimelere örnek olarak gösterilebilir.

5.2.2.3 Özellik vektörün boyutunu azaltmak ve dur kelimeleri

[12] Bu konuda daha detaylı bilgiyi 2. Kısımda Biçimbilimsel Analiz başlığı altında bulabilirsiniz.

Öğrenme algoritmaları bir nevi genelleştirme yaparlar. Yakın ve birbirini tekrarlayan özellikler tek bir özelliğe indirgenerek modellenmektedir. Algoritmanın işlem zamanını düşürmek ve doğruluk performansını arttırmak için özellik vektörünün boyutu mümkün olduğunca küçük tutulmalıdır. Dilde çokça kullanılan bazı kelimeler ayırt edici bir etki barındırmayıp yaygın şekilde metin içerisinde kullanılmaktır. Bu tip kelimelere dur kelimeleri (stop words) denir. 've', 'ile', 'ama', 'ancak', vb gibi bağlaçlar; 'ben', 'bu', vb gibi zamirler; dur kelimelerine örnek olarak verilebilir. Dur kelimeleri kullanılan dil ve veri kümesine göre farklılık gösterebilmektedir. Literatürde bazı dillere ait dur kelimelerini çıkartmak amacıyla yapılmış çalışmalar mevcuttur[13]. Dur kelimelerinin filtrelenmesi özellik vektörünün boyutunu düşürmekte, aynı zamanda performansı olumlu yönde etkilemektedir.

5.2.3 Eğitim verisinin belirlenmesi

Eğitim verisi yapay zekâ, yapay öğrenme, istatistiksel sistemler, genetik programlama gibi nesneler arasındaki ilişkiyi kullanarak model oluşturan tüm alanlarda sistemin eğitilmesi için ihtiyaç duyulan veri kümesidir. Gözetimli öğrenmede eğitim verisi her bir özellik vektörüne karşılık gelen etiket değerini de içerir. Tercih edilen değerlendirme yöntemine göre işaretli veri kümesi eğitim verisi ve test verisi olarak ikiye ayrılır. Literatürdeki pek çok çalışmada işaretlenmiş veri kümesinin %70'i veya 2/ 3'ü eğitim verisi olarak kullanılmaktadır.

5.2.4 Öğrenme algoritmasının seçimi

Uygun öğrenme algoritması problemin cinsine ve veri kümesine göre değişiklik göstermektedir. Örneğin çekirdek makineleri ve yapay sinir ağları en mükemmel sonuç için büyük öğrenme verisine ihtiyaç duyarken Naïve Bayes yöntemi daha az öğrenme verisine ihtiyaç duyar. Naïve Bayes öğrenme ve sınıflandırma aşamasında az bir depolama alanı kullanırken, kNN algoritması büyük depolama alanı kullanır[14].

[13]http://www.ranks.nl/stopwords adresinde bazı dillere ait dur kelimeleri mevcuttur.
[14]Algoritmalar arasındaki farkları görmek için Çizelge 3.2'ye bakabilirsiniz.

5.2.5 Başarımı değerlendirme yöntemleri

Bir sınıflandırıcının başarımı belirlenen ölçütler kullanılarak hesaplanır. Literatürde en sık kullanılan başarım ölçütleri; tutturma (precision), bulma (recall), doğruluk (accuracy) oranlarıdır.

Tutturma oranı (precision) aslen pozitif değer olup sınıflandırıcı tarafından pozitif olarak tespit edilenlerin sayısının tüm pozitif olarak tespit edilenlerin sayısına oranıdır **(5.3)**.

$$S_{hassalık\ oranı} = \frac{TP}{TP + FP} \tag{5.3}$$

Bulma oranı (recall) aslen pozitif değer olup sınıflandırıcı tarafından pozitif olarak tespit edilenlerin sayısının gerçekte pozitif olanlara oranıdır **(5.4)**.

$$S_{geri\ çağırma\ oranı} = \frac{TP}{TP + FN} \tag{5.4}$$

Doğruluğu (accuracy) matematiksel olarak doğru tahminlerin tüm tahminlere oranı şeklinde hesaplanır **(5.5)**.

$$S_{doğruluk\ oranı} = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.5}$$

Bu formüllerde; S, bir sınıflandırıcı; TP, aslen pozitif değer olup sınıflandırıcı tarafından pozitif olarak tahmin edilen; FP, aslen negatif değer olup sınıflandırıcı tarafından pozitif olarak tahmin edilen; TN, aslen negatif değer olup sınıflandırıcı tarafından negatif olarak tahmin edilen; FN, aslen pozitif olup sınıflandırıcı tarafından negatif olarak tahmin edilen bulgu sayısını ifade etmektedir.

Sınıflandırma başarımını daha iyi resmetmek için tutturma oranı (precision) ve bulma oranı (recall) başarım ölçütleri baz alınarak F_1 Skor (F_1 Score) ve G Skor (G Score) yöntemleri geliştirilmiş (Ikonomakis ve diğerleri, 2005). F_1 Skor tutturma ve bulma oranlarının harmonik ortalamasını alırken **(5.6)** ve G Skor (G Score) geometrik ortalamasını almaktadır **(5.7)**.

$$F_1 Skor = 2.\frac{S_{hassalık\ oranı}.S_{geri\ çağırma\ oranı}}{S_{hassalık\ oranı} + S_{geri\ çağırma\ oranı}} \quad \textbf{(5.6)}$$

$$G\ Skor = \sqrt{S_{hassalık\ oranı}.S_{geri\ çağırma\ oranı}} \quad \textbf{(5.7)}$$

Sınıflandırıcıların doğruluğunu hesaplamak için literatürde pek çok teknik önerilmiştir. Burada sık karşılaşılan üç değerlendirme türüne değinilecektir. Bunlar şu şekildedir ;

1. İşaretlenmiş veri kümesinin büyük bir kısmı eğitim verisi olarak kullanılırken, kalan diğer kısmı test verisi olarak kullanılmaktadır. Böylelikle daha önceden görülmeyen işaretlenmiş veri üzerindeki sınıflandırıcı tahminleriyle işaretlenen etiket kıyaslanarak sınıflandırıcının doğruluğu tahmin edilmeye çalışılmaktadır. Diğer yaklaşımlara göre daha düşük maliyetlidir. Bu yüzden yaygın şekilde kullanılır.

2. Diğer bir teknik çapraz doğrulama (cross-validation) olarak bilinir. Bir deneyin bağımsız koşullarda yinelenerek geçerliliğinin sınanması mantığına dayanmaktadır. İşaretlenmiş veri kümesi n adet eşit alt kümeye bölünür. Her bir alt küme k için, k kümesi haricindeki diğer n-1 adet alt küme birleştirilerek sınıflandırıcıyı eğitmek için kullanılır ve k. veri kümesiyle kıyaslanarak hata oranı (error rate) hesaplanır. Nihayetinde tüm hata oranlarının ortalaması alınarak sınıflandırıcının hata oranını tahmin edilir.

3. Leave-one-out çapraz doğrulama (LOOCV) yaklaşımındaysa her bir alt küme yalnız bir örnek içerecek şekilde bölünür. Çapraz doğrulama ile aynı yol izlenir. Çok maliyetli bir doğrulama tekniğidir. Hassas doğruluk ölçümleri için kullanılır.

5.2.6 Başarım oranının yetersiz olmasının sebebleri

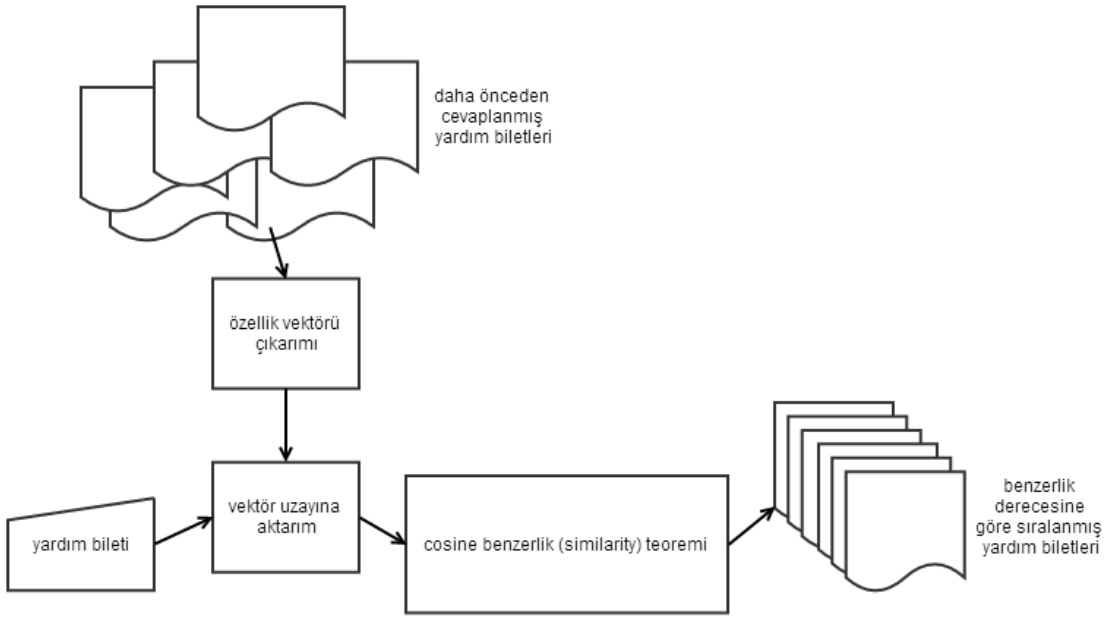
Genel olarak sınıflandırıcının başarım oranının yetersiz olmasının sebepleri aşağıdaki gibi sıralanmıştır (Kotsiantis, 2007).

- Sınıflandırma için yararlı, bilgi içeren özellikler yeteri kadar kullanılmamış olabilir.
- Daha geniş bir eğitim kümesine ihtiyaç olabilir.
- Problem çok karmaşık ve özellik vektör boyutu aşırı yüksek olabilir.

- Sınıflandırıcıya ait parametreler uygun ayarlanmamış olabilir.
- Sınıflandırma için seçilen algoritma uygun olmamış olabilir.
- Kullanılan veri kümesi uygun olmamış olabilir.

5.3 Yardım Biletlerinin Otomatik Cevaplandırılması

Çoğu zaman aynı veya benzer yardım biletleri ile karşılaşmak mümkündür. Önerilen sistem kaynak tüketimini minimum düzeye indirmeyi ve kullanıcı memnuniyetini azami dereceye çıkarmayı amaçlamaktadır. Bu sebeble önerilen sistemde yardım biletlerinin benzerlik (similarity) ve metin işleme algoritmaları kullanılarak otomatik olarak cevaplanması hedeflenmiştir.



Şekil 5.3 : Yardım biletlerinin benzerlik oranlarına göre sıralanması

Şekil 5.3' te daha önceden cevaplanan yardım biletlerinin yeni gelen yardım bileti ile benzerlik oranlarına göre sıralanmasını sağlayan sistem mimarisi verilmiştir. Kullanıcıdan gelen yardım bileti, daha önce cevaplanmış yardım biletleri baz alınarak elde edilmiş özellik vektörü ve uygun terim ağırlıklandırma metodu kullanılarak vektör uzayına aktarılılır. Vektör uzayında her bir bilet vektör olarak tanımlanmaktadır. Cosine bezerlik ölçütü **(5.8)** ile daha önceden cevaplanmış veri kümesi içerisindeki yardım biletlerine ait vektörlerin, yeni gelen yardım biletine ait vektöre olan uzaklıkları hesaplanır ve yardım biletleri benzerlik oranlarına göre sıralandırılır. Sıralanmış yardım biletleri içerisinden daha

önceden belirlenmiş eşik değerini geçen ilk n yardım biletine verilmiş cevap, kullanıcıya çözüm olarak sunulur.

Aşağıdaki denklemde a_i ve b_i vektör uzayında bulunan iki ayrı yardım biletini betimlemektedir. Denklemden elde edilecelek sonuç; iki yardım biletine ait vektörler eğer aynı ise 1, eğer birbirine dik ise 0 olacaktır.

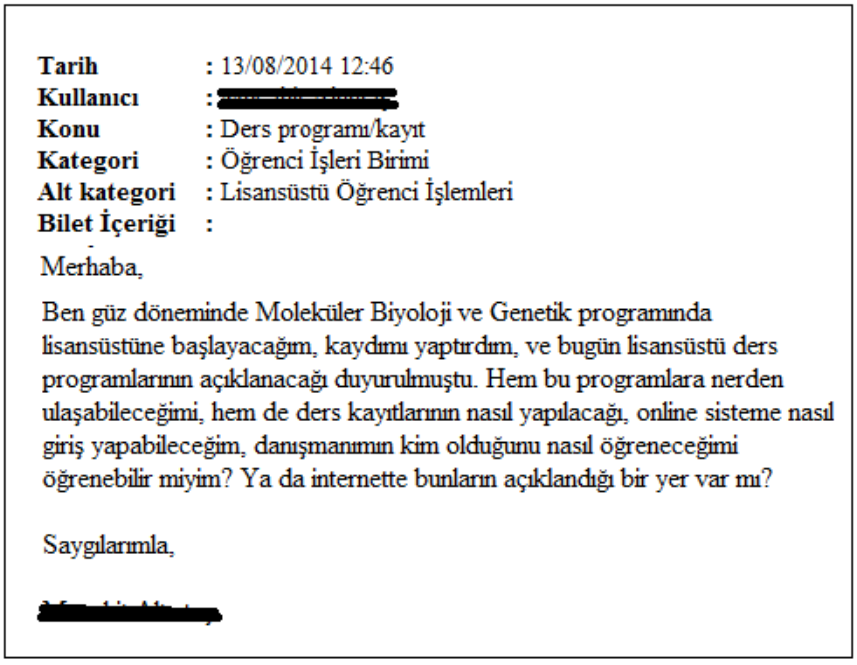
$$similarity(a_i, b_i) = \sum_{i=1}^{n} a_i . b_i \Big/ \left(\sqrt{\sum_{i=1}^{n} a_i^2} . \sqrt{\sum_{i=1}^{n} b_i^2} \right) \quad \textbf{(5.8)}$$

6. UYGULAMANIN GERÇEKLEŞTİRİLMESİ VE SONUÇLAR

Önerdiğimiz sistem İTÜ Yardım Portalı verileri kullanılarak uygulamaya dönüştürülmüştür. Çalışmanın bu kısmında uygulama aşamalarından bahsedilmiş ve elde edilen sonuçlar incelenmiştir.

6.1 Veri Kümesi

Bu çalışmada İTÜ Yardım Portalı'na ait, Türkçe doğal dili ile yazılmış yaklaşık on bin yardım bileti içeren veri kümesi kullanılmıştır. Yardım biletlerinin her biri tarih, kullanıcı bilgisi, ilgili bölüm, ilgili problem tipi, yardım konusu ve bilet içeriğini barındırmaktadır. Bu veri kümesindeki ilgili bölüm bilgisi önerilen sistemdeki kategoriye, ilgili problem tipi ise alt kategoriye karşılık gelmektedir. Şekil 6.1 veri kümesine ait tipik bir yardım bileti örneğini göstermektedir.



Tarih : 13/08/2014 12:46
Kullanıcı : [redacted]
Konu : Ders programı/kayıt
Kategori : Öğrenci İşleri Birimi
Alt kategori : Lisansüstü Öğrenci İşlemleri
Bilet İçeriği :

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,

[redacted]

Şekil 6.1 : Kullanılan veri kümesine ait tipik bir yardım bileti örneği

Çalışmadaki amaç, daha önce de belirtildiği gibi son kullanıcıdan gelen yardım biletlerinin konu ve içeriğine göre sınıflandırılıp ilgili destek personeli veya birimine

aktarılmak üzere sınıflandırılmasıdır. Sınıflandırma işlemi için bilete ait konu başlığı ve bilet içeriği kullanılmaktadır. Gönderici bilgisi ve gönderilme tarihi gibi bilgiler amacımıza yönelik uygun bilgi içermediği için göz ardı edilmiştir.

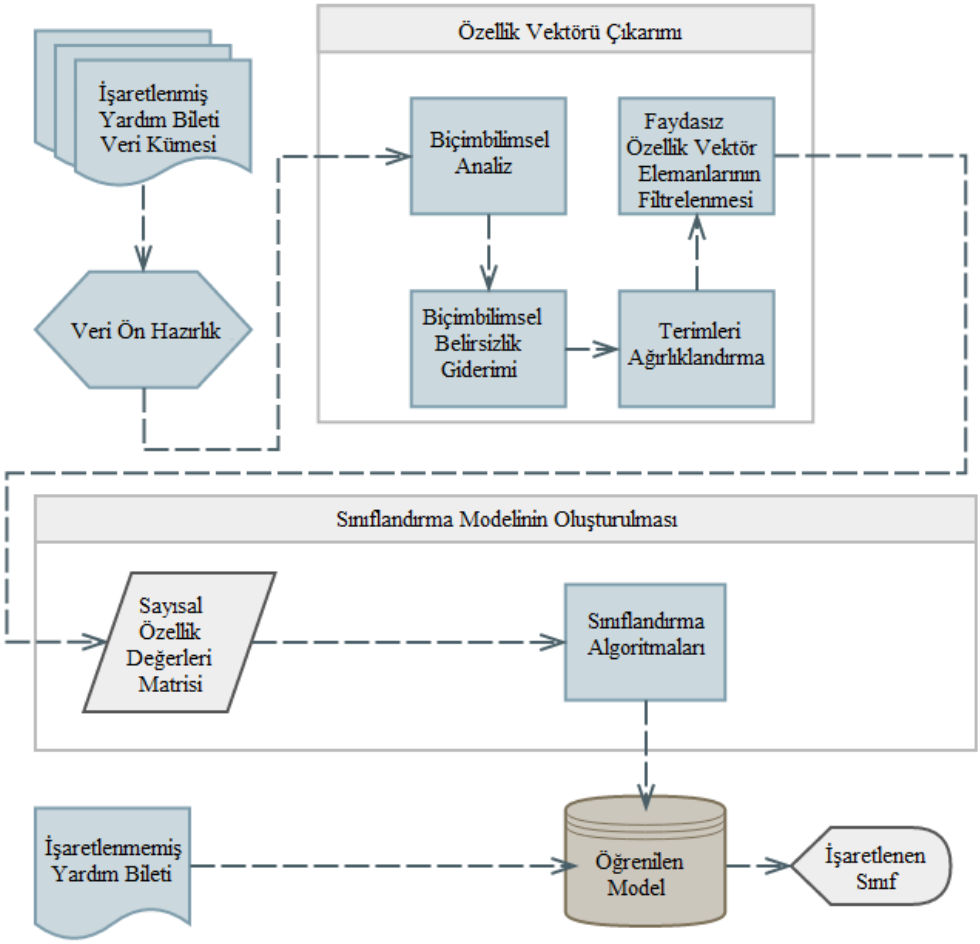
Çizelge 6.1 : Uygulamada kullanılan veri kümesine ait yardım biletlerinin kategori ve alt kategorine göre dağılımı

Kategoriler	**Alt kategoriler**	**Alt kategoriye ait bilet sayısı**	**Kategoriye ait bilet sayısı**
Bilgi İşlem Birimi	E-postalar	842	3380
	Kampüs kart seçimi	423	
	Geliştirilen servisler	1232	
	Diğer uygulama ve servisler	712	
	Diğerleri	671	
Öğrenci İşleri Birimi	Çift anadal öğrenci işlemleri	405	6586
	Lisans ders programları	285	
	Lisansüstü ders programları	829	
	Sisteme giriş problemleri	311	
	Mezuniyet işlemleri	202	
	Lisansüstü öğrenci işlemleri	303	
	PIN İşlemleri	3122	
	Katkı payı ücretleri	224	
	Diğerleri	905	
Sağlık, Kültür ve Spor Birimi	Konaklama	20	130
	Beslenme	30	
	Bilimsel aktiviteleri destekleme	35	
	Diğerleri	45	
Yurt ve Burs Birimi	Öneri ve tavsiyeler	68	113
	Şikayetler	44	
TOPLAM			10209

Çizelge 6.1 kullanılan veri kümesine ait biletlerinin kategori ve alt kategori bazında dağılımını göstermektedir.

6.2 Ön Hazırlık İşlemleri

Yardım portalı web tabanlı bir uygulama olduğu için gelen yardım biletlerinde html etiketler bulunabilmektedir. Ön işleme aşamasında verimiz html etiketlerden arındırılmıştır. Ayrıca sınıflandırma işlemine katkı sağlamadığı için bilet içeriğinde yer alan nümerik ifadeler kaldırılmıştır.



Şekil 6.2 : Sınıflandırma modelinin oluşturulamasında izlenen işlem akışı

6.3 Özellik Vektörü Çıkarımı

Özellik vektörü oluşturulurken en sık karşılaşılan problemlerin başında veri seyrekliği, gereksiz özellik vektörü elemanları ve özellik vektör boyutunun hayli yüksek olması gelmektedir. Devam eden kısımda karşılaşılan bu problemlere çözümler ve çalışmada gerçekleştirilen adımlardan bahsedilecektir. Şekil 6.2' de verilen sınıflandırma işlemi için izlenen akış şemasında özellik vektörü çıkarımı için uygulama basamakları görülebilir.

6.3.1 Muhtemel Kelime Köklerinin Bulunması

Türkçe gibi sondan eklemeli dillerde kelime, köküne aldığı ek ile metin içerisinde birden fazla formda bulunabilmektedir. Bu durum veri seyrekliği problemine neden

olmaktadır. Aşağıda kullandığımız veri kümesi içinde sık geçen ‘kart’ kelimesinin muhtemel çekim ve yapım ekleriyle gelebilecek yüzey formlarının morfolojik analiz sonuçları verilmiştir.

kart:

kart+Adj

kart+Noun+A3sg+Pnon+Nom

kartım:

kart+Noun+A3sg+Pnon+Nom^DB+Verb+Zero+Pres+A1sg

kart+Noun+A3sg+P1sg+Nom

kartımı:

kart+Noun+A3sg+P1sg+Acc

karta:

kart+Noun+A3sg+Pnon+Dat

kartı:

kart+Noun+A3sg+Pnon+Acc

kart+Noun+A3sg+P3sg+Nom

kartlar:

kart+Noun+A3pl+Pnon+Nom

kart+Noun+A3sg+Pnon+Nom^DB+Verb+Zero+Pres+A3pl

kartın:

kart+Noun+A3sg+Pnon+Gen

kart+Noun+A3sg+P2sg+Nom

kartlı:

kartlı+Noun+A3sg+Pnon+Nom

kartına:

kart+Noun+A3sg+P2sg+Dat

kart+Noun+A3sg+P3sg+Dat

Çalışmada veri seyrekliği problemiyle başa çıkmak için veri kümesi içerisinde geçen kelimelerin morfolojik analiz ile muhtemel kök ve ekleri bulunmuş, sonrasında; morfolojik analiz sonuçları, yapısal belirsizlik gidericiden geçirilerek en olası kök özellik vektörü elemanı olarak kabul edilmiştir. Bu yöntemin daha basit ve külfetsiz olan ilk n karakter yöntemine göre özellik vektörü elaman sayısını daha iyi indirgediği görülmüştür. Çizelge 6.2 'deki sonuçlar aynı veri kümesi üzerinde, tf-idf terim ağırlıklandırma yöntemi kullanılarak gerçekleştirilmiştir.

Çizelge 6.2 : Veri seyrekliği problemine yönelik yaklaşımların başarımları

Özellik (Attributes)	Yöntem	Sınıflandırma Doğruluk Başarımı
500	Biçimbilimsel Analiz + Belirsizlik Giderimi	%76.92
501	İlk 3 karakter	%68.00
859	İlk 6 karakter	%67.30
1024	İlk 9 karakter	%65.38

Kelime köklerinin morfolojik analizi için ITU Turkish NLP Pipeline[15] aracı kullanılmıştır (Eryiğit, 2014). Her bir kelime için biçimbilisel analiz gerçekleştirilmiştir.

6.3.2 Biçimbilimsel Belirsizliğin Giderilmesi

Biçimbilimsel çözümlemesi yapılan kelimeler belirsizlik içermektedirler. Kelimenin anlamı cümle içerisinde geçen diğer kelimelerle ilişkilidir. Bu nedenle belirsizlik giderimi cümleler düzeyinde gerçekleşmektedir. Çalışmada elde morfolojik analiz sonuçları cümleler halinde belirsizlik gidericiye verilmiş ve cümle içerisindeki her kelimeye ait en olası morfolojik çözümleme cümleler halinde alınmıştır. Bu işlem için yine ITU Turkish NLP Pipeline kullanılmıştır. Biçimbilimsel belirsizlik giderimi işleminden sonra her bir kelimenin kökü özellik vektörü elemanı olarak tanımlanmıştır.

6.3.3 Terim Ağırlıklandırma

Bu çalışmada, literatürde sık kullanılan ağırlıklandırma yöntemlerinin sınıflandırma başarımlarımı üzerindeki etkisini inceleyebilmek için ikili terim ağırlıklandırma (boolean term weighting), terim sıklığı (term frequency) ve terim sıklığı – ters

[15] Detaylı bilgiye http://tools.nlp.itu.edu.tr/ adresinden ulaşabilirsiniz.

doküman sıklığı (term frequency – inverse document frequency) metotları farklı sınıflandırma algoritmaları ile birlikte sınanmıştır. Elde edilen sonuçlar, ağırlıklandırma metotlarının sınıflandırma başarımında önemli rol oynadığını göstermiştir.

6.3.4 Faydasız Özellik Vektörü Elemanlarının Filtrelenmesi

Uygulamamızda veri kümemizin genişliği ve geçen kelime sayısının çeşitliliğiyle bağlantılı olarak özellik vektör boyutumuz hayli yüksek olmuştur. Özellik vektöründeki bazı vektör elemanları sınıflandırma işlemi için fark edilir düzeyde ayırt edici bir rol oynamamaktadır. Çalışmamızda belirli eşik değeri altında idf değerine sahip özellik vektörü elemanları, sıradan kelimeler (stop words) olarak kabul edilmiş ve filtrelenerek özellik vektör boyutu mümkün olduğunca düşürülmüştür. Bunun yanında noktama işaretleri ve cümle içerisinde, konudan bağımsız olarak yaygın şekilde kullanılan bağlaçlar sınıflandırma için ayırt ediciliğe sahip olmadıkları için dur kelimesi olarak tanımlanmışlar ve özellik vektörü elamanları içerisinden çıkarılmışlardır.

Sıradan kelimeleri tespit etmek için kullandığımız yöntem kullanıcı tarafından yanlış yazılmış kelimelerden, özel isimlerden ve yabancı kelimelerden kurtulmamıza da olanak sağlamaktadır. Uyguladığımız yöntem ile veri kümesinde yakalanan bazı sıradan kelimeler Çizelge 6.3' de verilmiştir.

Çizelge 6.3 : Veri kümesinde yakalanan sıradan kelimelerden bazıları

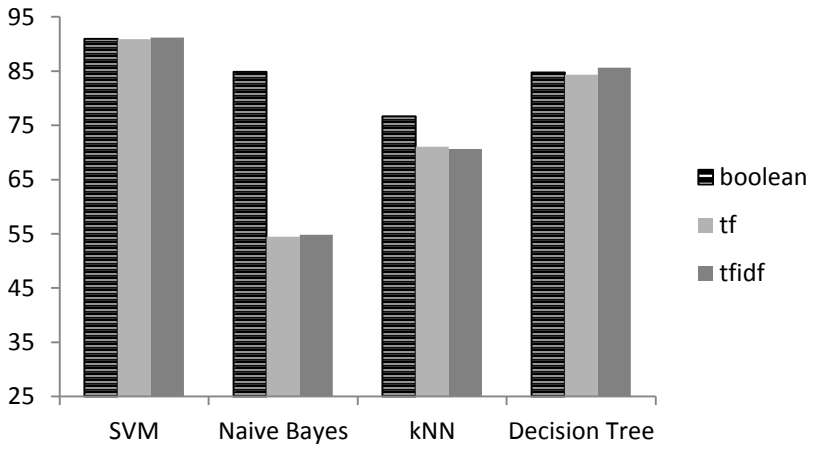
Bazı Sıradan Kelimeler	
start	main
cavlak	imdi
gültekin	gürhan
logos	ercan
meblağ	bani
sezer	hafize

6.4 Sınıflandırma Modellerinin Oluşturulması

Kullandığımız veri kümesi, dört faklı kategori ve her bir kategoriye ait çeşitli sayılarda alt kategori içermektedir. Önerdiğimiz yöntem kullanıcıdan gelen yardım biletinin alt kategorisini belirlemek için bileti iki farklı sınıflandırma işlemine tabi tutmaktadır. Bu sınıflandırma işlemlerinden ilki biletin kategorisini belirlemekte, ikincisi ise biletin

belirlenen kategori altındaki hangi alt kategoriye dâhil olduğunu tespit etmektedir. Bu amaçla her bir kategori için kategoriye ait alt kategorileri sınıflandırmak amacıyla bağımsız sınıflandırma modelleri oluşturulmuştur. Alt kategori sınıflandırıcılardan her bir model kendine ait veri kümesine sahiptir. Çalışmada açık kaynak veri madenciliği kütüphanesi WEKA kullanılmıştır. Uygulama java programlama dili ile yazılmıştır.

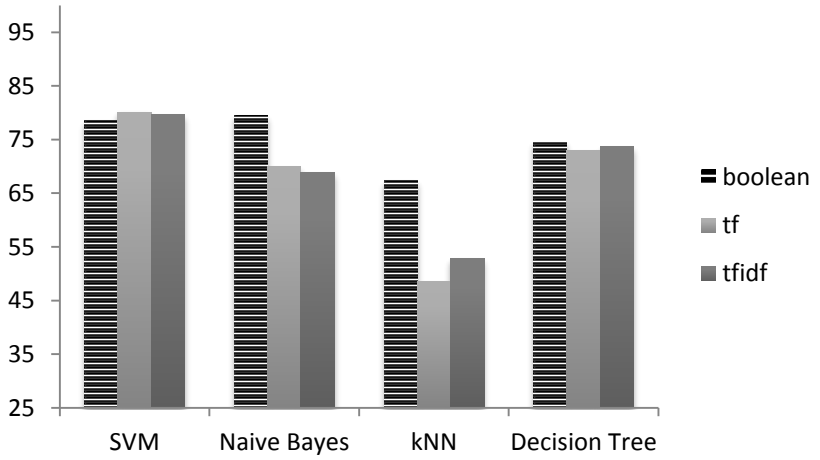
Bu çalışmada sınıflandırma performanslarını karşılaştırmak amacıyla dört farklı gözetimli yapay öğrenme algoritması kullanılmıştır. Bu algoritmalar sırasıyla; çok çekirdekli destek vektör makinaları (SVM), istatistiksel tabanlı Naïve Bayes, en yakın k komşu (kNN) ve karar ağaçları (decision tree) sınıflandırma algoritmalarıdır. Uygulamada, kNN sınıflandırma algoritması için k değeri 1 (bir) olarak alınmıştır. Karar ağacı algoritması budanmamış (unpruned) olarak uygulanmıştır. Sınıflandırma başarımlarını ölçmek için on kümeli çapraz doğrulama (ten fold cross validation) metodu kullanılmıştır. Farklı terim ağırlıklandırma yöntemleri kullanılarak elde edilen sonuçlar aşağıda verilmiştir. (Şekil 6.3 – 6.7)



Şekil 6.3 : Kategori sınıflandırmada doğruluk performansları

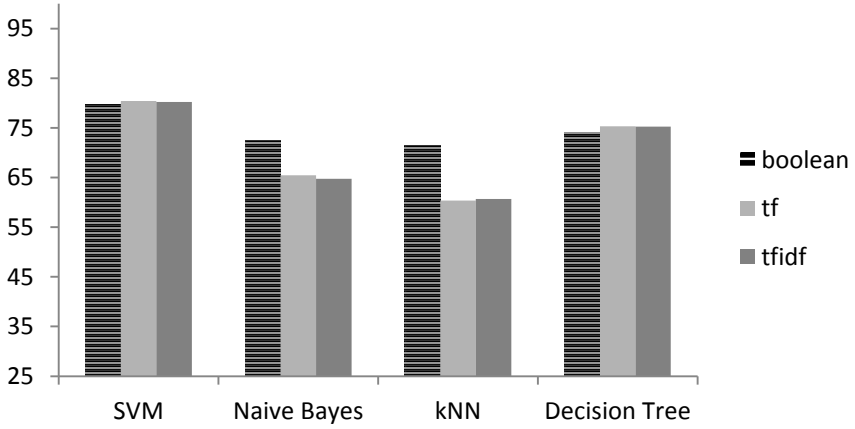
Destek vektör makinaları (SVM'ler) ayırtaç tabanlı sınıflandırıcılar olduklarından yalnızca sınıra yakın örneklerle ilgilenirler diğer örnekleri göz ardı ederler. Dolayısıyla sınıflandırıcının karmaşıklığı veri kümesinin boyutuna değil, destek vektör sayısına bağlıdır. Bu yüzden büyük veri içeren sınıflandırma problemleri için en uygun sınıflandırma yöntemidir. Hem doğrusal olarak ayrılabilen, hem de doğrusal olarak ayrılamayan veriler üzerinde etkilidirler. SVM'ler asgari risk esasını gözetmektedirler. Bu yüzden genel maksimumu bulma eğilimindedirler. Yapay sinir ağları vb diğer sınıflandırıcıların aksine yerel maksimumu bulduktan sonra öğrenme işlemi

tamamlanmaz. Ayrıca çekirdek makineleri parametre almayan (nonparametric) sınıflandırıcılardır. Bu yüzden öğrenme adımı, ilk değer, durma noktası gibi parametrelerin tanımlanmasına gerek yoktur.



Şekil 6.4 : Bilgi İşlem Birimi (kategorisi) altındaki alt kategorilerin sınıflandırılmasında doğruluk performansları

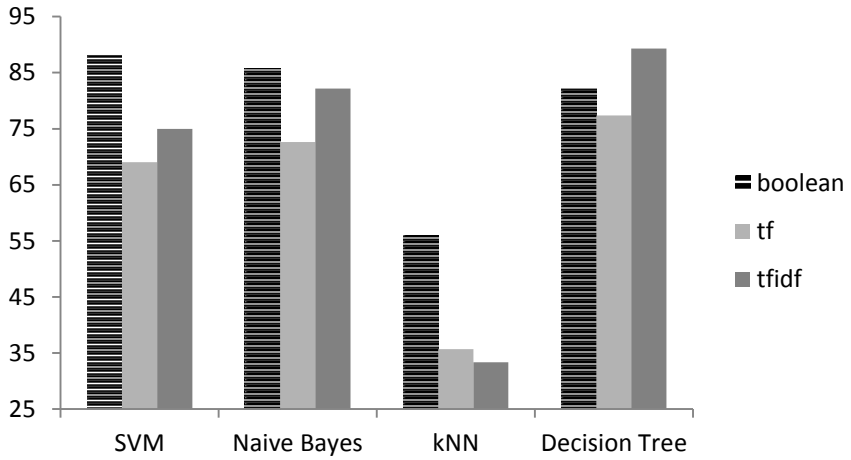
Naive Bayes algoritması Bayes teoremine dayanan istatistiksel tabanlı bir sınıflandırma tekniğidir. Tüm olasılıkları barındırmayan az sayıda öğrenme verisi içeren sınıflandırma problemlerinde diğer sınıflandırma algoritmalarına göre oldukça iyi başarım gösterir. Ayrıca Naive Bayes örneklerle değil özelliklerle ilgilenir. Bu yüzden diğer algoritmalara kıyasla öğrenme hızı daha yüksektir[16].



Şekil 6.5 : Öğrenci İşleri Birimi (kategorisi) altındaki alt kategorilerin sınıflandırılmasında doğruluk performansları

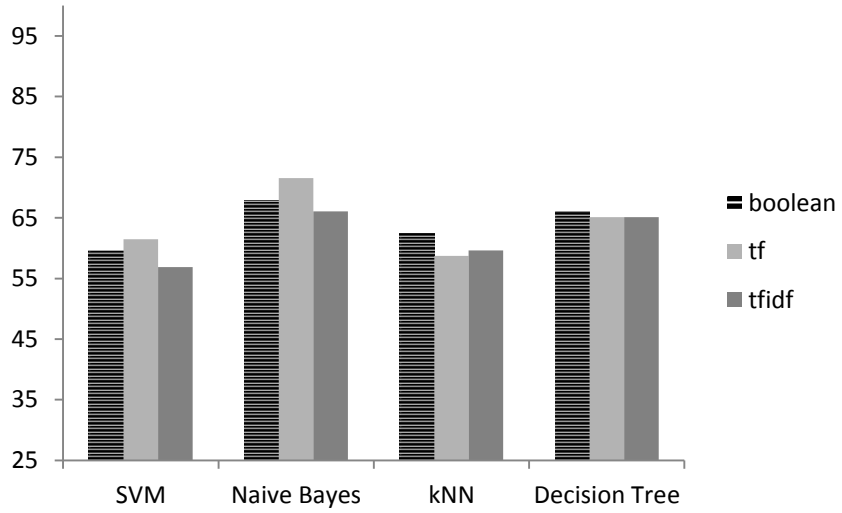
[16] Çalışmada sınanan algoritmalarının öğrenme hızları için Çizelge 6.3' e bakınız.

En yakın k komşu algoritması örnek tabanlı bir sınıflandırma algoritmasıdır. Verilerin dağılımı hakkında çok az ya da hiç bir bilginin olmadığı problemlerde genellikle bu algoritma tercih edilir. Ayrıca, bu algoritma sınıflandırma için en yakın k komşu örneği değerlendirdiği için (k>2 olmak şartıyla) gürültülü veriden daha az etkilenir.



Şekil 6.6 : Sağlık, Kültür ve Spor Birimi (kategorisi) altındaki alt kategorilerin sınıflandırılmasında doğruluk performansları

Karar ağaçları basit ve yaygın olarak kullanılan sınıflandırma tekniğidir. Bu algoritma yukardan aşağıya her düğüm noktasında, geride kalan verileri en uygun şekilde bölerek ikili ağaç yapısı oluşturur. Karar ağacının boyutunu azaltmak ve doğruluğunu artırmak amacıyla, sınıflandırma için zayıf bilgi içeren örnekler, budama işlemi ile öğrenme verisi içerisinden çıkarılırlar.



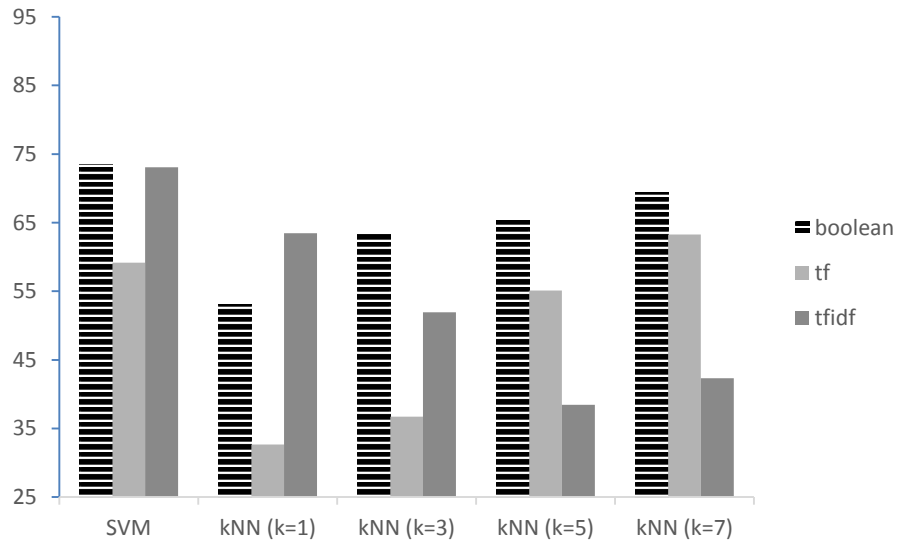
Şekil 6.7 : Yurt ve Burslar Birimi (kategorisi) altındaki alt kategorilerin sınıflandırılmasında doğruluk performansları

Çalışmada sınanan algoritmaların 0-100 arasında normalize edilmiş öğrenme hızları Çizelge 6.4'te verilmiştir[17].

Çizelge 6.4 : Sınıflandırma algoritmalarının öğrenme hızları

Sınıflandırma Algoritması	Normalize Edilmiş Öğrenme Hızı
SVM	3.98
Naive Bayes	90.33
kNN	27.47
Decision Tree	3.44

Sınıflandırma algoritmalarının doğruluk başarımı kullanılan veri kümesi ve terim ağırlıklandırma (term weighting) yöntemine göre değişiklik göstermektedir. Genel olarak SVM büyük veri kümesine sahip kategorilerde en iyi başarımı sağlarken, veri kümesinin daha küçük olduğu durumlarda (Yurt ve Burs Birimi alt kategorilerinde olduğu gibi) Naive Bayes'in gerisinde kalmaktadır. Karar ağaçları Naive Bayes ve kNN algoritmalarına nazaran daha iyi bir başarım sergilemiştir. En yakın k komşu algoritmasının başarımının düşük olmasında, k değerinin 1 (bir) alınması etkilidir. Şekil 6.8'de farklı k değerleri ile elde edilen başarımlar gösterilmiştir.



Şekil 6.8 : kNN algoritmasının değişen k değerlerine göre başarımları ve aynı veri kümesi üzerinde SVM başarımı

[17] Eğitim işlemi; 3800 adet örnek ve 1946 farklı özellik barındıran veri kümesi üzerinde Intel Core i7 2670QM 2.20 GHz CPU ile gerçekleştirilmiş.

Öğrenme hızları açısından değerlendirildiklerinde; istatistiksel öğrenme algoritması, Naive Bayes en iyi performansı sergilerken, karar ağaçları ve SVM algoritmalarını gerisinde bırakmıştır.

6.5 Otomatik Sınıflandırıcı Başarımının Yeterliliği

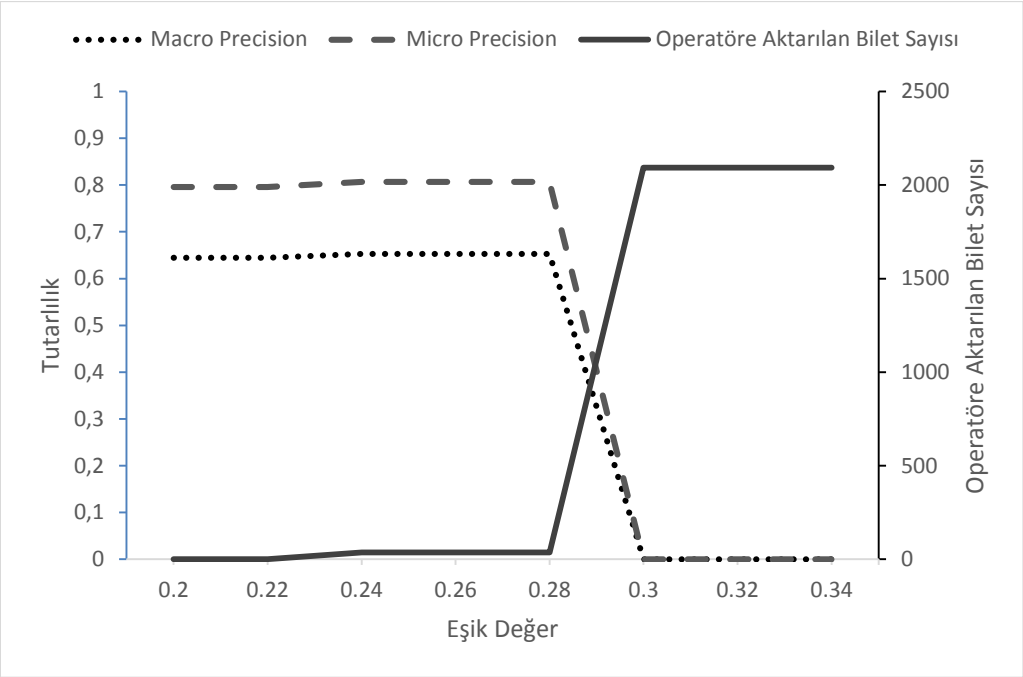
Önerdiğimiz sistemde, otomatik sınıflandırıcı tarafından yapılan sınıflandırmaların güven (confidence) yüzdelerine bakılarak, belirlenen eşik değerin altında kalan biletler yeniden sınıflandırılmak üzere operatöre gönderilmektedir.

Çizelge 6.5 : Eşik değere göre operatöre aktarılan, doğru sınıflandırılan ve yanlış sınıflandırılan biletlerin sayısı

Eşik Değer	Operatöre Aktarılan Bilet Sayısı	Doğru Sınıflandırılan Bilet Sayısı	Yanlış Sınıflandırılan Bilet Sayısı
0.2	0	1686	404
0.22	0	1686	404
0.24	37	1682	374
0.26	37	1682	374
0.28	37	1682	374
0.3	2093	0	0
0.32	2093	0	0
0.34	2093	0	0

Çizelge 6.5’ te dört sınıf 2093 örnek barındıran veri kümesinin kullanıldığı bir SVM sınıflandırıcıya ait tahminlerin güven yüzdeleri göz önünde bulundurularak, değişen eşik değerlerine göre operatöre aktarılan bilet sayısı, sınıflandırıcı tarafından doğru sınıflandırılan bilet sayısı ve sınıflandırıcı tarafından yanlış sınıflandırılan bilet sayısı verilmiştir.

Şekil 6.9’da operatöre aktarılan bilet sayısı, makro tutturma ve mikro tutturma oranları grafiksel şekilde gösterilmiştir. Eşik değer arttırıldıkça operatöre aktarılan bilet sayısı artmakta, yanlış sınıflandırılan bilet sayısı azalmaktadır. Eşik değer 0.3 değerine çekildiğinde, otomatik sınıflandırıcının tahminlerinin hiç birisi istenen eşik değerini geçememekte ve yardım biletlerinin hepsi sınıflandırılmak üzere operatöre gönderilmektedir. Tutturma oranları ve operatöre gönderilen bilet sayısı göz önünde bulundurularak en uygun eşik değerin 0.22 ile 0.28 arasında olduğu görülmektedir.



Şekil 6.9 : Uygulanan eşik değerlerine göre operatöre aktarılan bilet sayısı, makro tutturma (macro precision) ve mikro tutturma (micro precision) oranlarındaki değişim

6.6 Yardım Biletinin Otomatik Cevaplanması

Önerilen sistem, gelen yardım biletini daha önceden cevaplanan yardım biletleri içerisinden yakalayıp otomatik olarak cevaplamayı amaçlamaktadır. Gelen biletin daha önceden oluşturulan biletler ile arasındaki benzerliği ölçmek için cosine benzerlik (similarity) ölçütü kullanılmıştır. Bu ölçüt hesaplanırken, ciddi işlem yükü oluşmaktadır. Bu sebeple bu çalışmada, daha önceden cevaplanan biletler yerine İTÜ Kullanıcı Destek Sistemi'ne ait sık sorulan sorular kullanılmıştır. Bu sayede işlem yükünden bir miktar kurtulmak istenmiştir. Ancak, kullandığımız sık sorulan sorular sınırlı sayıda ve kullanıcıdan gelen biletler ile yeterli düzeyde örtüşmediği için, yakalanan benzer bilet sayısı az olmuştur. Benzerlik oranları, eşik değer olarak belirlediğimiz 0.5 'in altıda kalmışlardır. Eşik değerinin geçemeyen yardım biletleri için kullanıcıya herhangi otomatik cevap önerisi gitmemektedir.

Çizelge 6.6' da kullanıcıdan gelen yardım biletine Sık Sorulan Sorular (SSS) içerisinden tespit edilen benzer soru ve kullanıcıya önerilen cevap verilmiştir. Bu örnekte tespit edilen sık sorulan soru, 0.39 benzerlik oranı ile yakalanmıştır.

Çizelge 6.6 : Gelen yardım bileti, tespit edilen benzer SSS ve önerilen cevap

Gelen Bilet	*"evrak kayıt programını kullanamıyorum. yetkilendirme hatası veriyor. nasıl kayıt yapabilirim? iyi günler."*
Benzer Sık Sorulan Soru	*"evrak kayıt programında yetkilendirilmek için ne yapmalıyım?"*
Önerilen Cevap	*"ekp üniversitenin idari birimlerinde gelen-giden evrakların kaydını tutan ve raporlayan bir uygulamadır. yeni bir kullanıcının ekp'yi kullanabilmesi için kendi birim yöneticisi tarafından yetkilendirilmesiyle ilgili dilekçenin yazılması gereklidir."*

7. DEĞERLENDİRME VE ÖNERİLER

7.1 Değerlendirme

Günümüz rekabet koşullarında, organizasyonlar müşteri memnuniyetini azami derecede yüksek tutmayı amaçlamaktadırlar. Destek sistemleri, hizmet veya ürün alımı öncesi ve sonrasında bu amaca katkı sağlamakta önemli rol oynamaktadır. Bu sistemlerin verimliğini ve kullanılabilirliğini arttırmak, organizasyonlar için hayati önem arz etmektedir.

Destek sistemleri, kullanıcıdan gelen istek ve görüşleri yardım bileti şeklinde tutarak, ilgili destek personeli veya birimi tarafından cevaplanmasını sağlar. Bu işlemin hızlı ve etkin şekilde gerçekleştirilmesi, özellikle büyük organizasyonlar için kritik önemdedir. Bu çalışmada, destek sistemlerinin iyileştirmesine yönelik yeni bir destek sistem mimarisi önerilmiştir. Önerilen sistem mimarisi, İTÜ Destek Sistemi verileri kullanılarak uygulama haline getirilmiştir. Uygulamada elde edilen sonuçlar önerilen sistemin büyük organizasyonların ihtiyaçlarını karşılayacak etkinliğe sahip olduğunu göstermiştir.

Çalışmada, yapay zekâ algoritmaları ve doğal dil işleme yöntemleri kullanılarak yardım biletlerinin ilgili birime otomatik olarak aktarılması gerçekleştirilmiş, bunun yanı sıra tekrar eden yardım biletlerinin sistem tarafından otomatik cevaplanması sağlanmıştır. Önerilen sistem sayesinde;

- Yardım biletlerinin ilgili birime atama işlemi otomatikleştirilerek bu işlemin destek operatörleri üzerindeki yükü, azaltılmıştır. Bu sayede zaman alıcı ve hataya meyilli insan emeği gerektiren elle atama işlemi asgari düzeye indirgenmiştir.
- Atama işleminin otomatikleştirilmesiyle, kullanıcıdan gelen yardım biletleri daha kısa sürede cevaplanarak sağlanan servis kalitesi, dolayısıyla son-kullanıcı memnuniyeti arttırılmıştır.

- Yanlış atama işlemleri minimuma indirgenerek, değerli destek kaynaklarının daha etkin ve verimli şekilde kullanılmasına olanak sağlanmıştır. Ayrıca tekrar eden yardım biletlerinin tespit edilip otomatik cevaplandırılması, destek personelinin iş yükünü azaltmıştır.

7.2 Öneriler

Bu çalışmada önerdiğimiz sistemin verimliliğini arttırmak amacıyla, bazı iyileştirmeler yapılabilir. Bunlardan bazıları şu şekildedir;

- Yardım biletleri kendi içinde bir bütünlüğe sahiptir. Gerçekleştirilen uygulamada konu başlığı ve bilet içeriği analiz edilmektedir. Eğer sadece konu başlığı veya sadece mesaj içeriği analiz edilirse işlem zamanı düşürülebilir. Hatta ilk n kelime üzerinden analiz yapılabilir.

- Mevcut uygulamada özellik vektörü tüm terimleri içermektedir. Eğer sadece metin içerisindeki varlık isimleri kullanılırsa, özellik vektör boyutu büyük ölçüde küçültülebilir. Bu hem işlem süresini düşürür, hem de gürültüden bir miktar kurtulunacağı için doğruluk oranını yükseltir.

- Kullanıcıdan gelen yardım biletleri nadir de olsa yanlış yazılmış terimler içermektedir. Gerçekleştirdiğimiz uygulamada bu terimler ön işleme aşamasında tespit edilip, elimine edilmektedir. Bazen yanlış yazılan bu terimler bilet konusunu belirlemede önemli rol oynamaktadır. Bu yüzden veri ön işleme aşamasında yanlış yazılan kelimelerin düzgün forma getirilmesi uygulamanın doğruluğunun arttırılmasında katkı sağlayacaktır.

- Gerçekleştirilen uygulama kelimelerin biçimbilimsel analizlerini göz önünde bulundurarak işlem yapmaktadır. Eğer kelimeler anlamsal analiz düzeyinde değerlendirilirse sistem başarımı artacaktır.

KAYNAKLAR

Akkus, B. K., & Cakici, R. (2013). Categorization of Turkish News Documents with Morphological Analysis. In ACL (Student Research Workshop) (pp. 1-8).

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

Amasyalı, M. F.,& Diri, B. (2006). Automaticturkishtextcategorization in terms of author, genreandgender. In Natural Language Processingand Information Systems(pp. 221-226). Springer Berlin Heidelberg. ISO 690

Amasyalı, M. F., & Yıldırım, T. (2004). Otomatik haber metinleri sınıflandırma. SIU 2004, 224-226.

Borko, H., & Bernick, M. (1963). Automatic document classification. Journal of the ACM (JACM), 10(2), 151-162.

Boser, B. E.,Guyon, I. M., Vapnik, V. N. (1992). A training algorithm for optimal magrin classifiers, COLT '92 Proceedings of the fifth annual workshop on Computational learning theory 144-152.

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, June

Crawford, E. (2004) Phrases and Feature Selection in E-Mail Classification, 9th Australasian Document Computing Symposium.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Çıltık, A.,& Güngör, T. (2008). Time-efficientspam e-mail filteringusing n-gram models. PatternRecognitionLetters, 29(1), 19-33.

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). Customerinteraction2.0: Adopting social media as customer service channel. Journal of advances in information technology, 2(4), 222-233.

Güran, A., Akyokuş, S., Bayazıt, N. G., & Gürbüz, M. Z. (2009). Turkish text categorization using N-gram words. In Proceedings of the International

Symposium on Innovations in Intelligent Systems and Applications (INISTA 2009) (pp. 369-373).

Harris, Z. S. (1954). Distributional structure. Word.

Hsu, C., Chang, C., Lin, C. (2010). A Practical Guide to Support Vector Classification.

Ikonomokis, M.,Konsiantis, S. &Tampakas, V. (2005) Text Classification Using Machine Learning Techniques, WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 4, August 2005, pp. 966-974.

Joachims, T. (1998) Text Categorization with Suport Vector Machines: Learning with Many Relevant Features, ECML '98 Proceedings of the 10th European Conference on Machine Learning 137-142.

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.

Klimt, B. & Yang, Y. (2004).The Enron Corpus: A New Dataset for Email Classification Research, Machine Learning: ECML 2004,Lecture Notes in Computer Science Volume 3201, 2004, pp 217-226.

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

Lewis, D. D., & Ringuette, M. (1994, April). A comparison of two learning algorithms for text categorization. In Third annual symposium on document analysis and information retrieval (Vol. 33, pp. 81-93).

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development, 1(4), 309-317.

Manco, G.,Masciari, E., Ruffolo, M., Tagerelli, A., (2002). Towards An Adaptive Mail Classifier, Artificial Intelligence Workshop Su Apprendimento Automatico: Metodi Ed Applicazioni.

Manning, C. & Schütze, H. (1999) Foundations of Statistical Natural Language Processing, MIT Press. Cambridge.

Maron, M. E. (1961). Automatic indexing: an experimental inquiry. Journal of the ACM (JACM), 8(3), 404-417.

Medlock, B. W. (2008) Investigating classification for natural language processing task, This technical report is based on a dissertation submitted Technical reports published by the University of Cambridge ISSN 1476-2986.

Nabiyev, V. V. (2010). Yapay Zeka İnsan-Bilgisayar Etkileşimi, 3. Baskı, Ankara, Şubat.

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

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

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

Şahin, M., Sulubacak, U., Eryiğit, G. (2013) Redefinition of Turkish Morphology Using Flag Diacritics. In Proceedings of the Tenth Symposium on Natural Language Processing (SNLP-2013), Phuket, Thailand.

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

Toroman, Ç. (2011) Text Categorization And Ensemble Pruning In Turkish News Portals, A Thesis Submitted To The Department Of Computer Engineering And Graduate School Of Engineering And Science Of Bilkent University.

Torunoglu, D., Cakirman, E., Ganiz, M. C., Akyokus, S., & Gurbuz, M. Z. (2011, June). Analysis of preprocessing methods on classification of Turkish texts. In Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on (pp. 112-117). IEEE.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 433-460.

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).

Vapnik, V. ve Cortes C. (1995). Support-Vector Networks, Kluwer Academic Publishers, Boston. Manufactured in The Netherlands,Machine Leaming, 20, 273-297.

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. Ve 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/

Youn, S., & McLeod, D. (2007). A comparative study for email classification. In Advances and Innovations in Systems, Computing Sciences and Software Engineering (pp. 387-391). Springer Netherlands.

Url-1<*http://www.csie.ntu.edu.tw/~cjlin/libsvm/* >, alındığı tarih: 16.02.2014.

Url-3<*http://www.cs.sfu.ca/research/groups/NLL/1.html*>,alındığı tarih: 19.01.2014.

Url-4<*http://tools.nlp.itu.edu.tr/* >, alındığı tarih: 23.02.2014.

Url-5<*http://www.ranks.nl/stopwords/* >, alındığı tarih: 24.03.2014

ÖZGEÇMİŞ

Ad Soyad: **Mücahit ALTINTAŞ**

Doğum Yeri ve Tarihi: **Balıkesir, 1988**

E-Posta: **maltintas@itu.edu.tr**

Yayın ve Bildiriler:

Altintas, M., Tantug, C. A.,(2014) Machine Learning Based Ticket Classification In Issue Tracking System, AICS, Artificial Intelligence and Computer Science, Proceeding