Tıbbi Kayıtlara ICD-10 Hastalık Kodlarının Atanmasına Yardımcı Akıllı Bir Sistem

Nefise Meltem CEYLAN [a] Adil ALPKOÇAK [a], Afsun Ezel ESATOĞLU[b]

[a] *Bilgisayar Mühendisliği Bölümü, Dokuz Eylül Üniversitesi, Tınaztepe, İzmir*
[b] *Sağlık Bilimleri Fakültesi, Ankara Üniversitesi, Ankara*

An Intelligent System to Help on Assignment of ICD-10 Codes to Medical Records

Abstract: *In this study we present an intelligent system proposal to help on assigning ICD10 codes to medical records. System is based on classification of free-form texts fields for a given medical record, and uses Terrier information retrieval engine to handle unstructured text data. In this way, we propose a new effective and efficient text classification method, which is based on information retrieval techniques. To evaluate the new approach we compiled a new Turkish medical collection, which includes approximately 50K preprocessed and anonimized medical records taken from Ankara University Hospital information systems. We measured the performance of proposed system in terms accuracy, selectivity and sensitivity.*

Keywords: Turkish Medical Text Classification, Auto–assigning ICD-10 Codes, Information Retrieval, Terrier

Özet: *Bu çalışmada tıbbi kayıtlara ICD kodlarının atanmasına yardımcı olacak akıllı bir sistemin önerisi sunulmuştur. Sistem temel olarak verilen bir sağlık kaydında yer alan serbest formda yazılmış metin bilgilerinin sınıflandırmasına dayalıdır ve serbest formda yazılmış metin sahalarını işlemek için açık kaynak kodlu Terrier bilgi gerigetirim sistemini kullanır. Bu anlamda çalışmayı Türkçe tıbbı dokümanların sınıflandırması olarak da değerlendirmek mümkündür. Bu bağlamda, serbest formda yazılmış metinlerin sınıflandırılmasında bilgi gerigetirim tekniklerini kullanan etkin ve verimli bir sınıflandırma yaklaşımı öneriyoruz. Sistemin sınanması için Ankara Üniversitesi Hastanesi bilgi sisteminden yaklaşık 50 bin kayıt alınarak, önişlemden geçirip anonimize ederek, bir Türkçe tıbbi veri seti oluşturduk. Önerdiğimiz sistemin başarım performansları doğruluk, özgüllük ve duyarlılık ölçeklerine göre ölçülmüştür.*

Anahtar Kelimeler: Türkçe Tıbbi Metin Sınıflandırma, ICD-10 kodlarının otomatik atanması, Bilgi Gerigetirim, Terrier.

1. Giriş

Tıbbi Bilgi sistemleri birçok amaç için yapısal veri tutar fakat bunun yanı sıra yapısal olmayan verileri de hasta kayıtlarında daha detaylı bilgi içermesi açısından serbest metin olarak barındırırlar. Yapısal olmayan bu veriyi bilgi sistemleri açısından anlamlı

hale getirmek önemli ve ciddi bir süreçtir. Sağlık harcamalarının düşürülüp sağlık hizmeti kalitesinin arttırılması kritik önem arz eder. Bu bağlamda hastalık kodlarının hasta raporları ile ilişkilendirilmesi henüz çözümlenememiş bir problemdir. ICD-10 klinik raporların kodlanmasında bir standart sunmaktadır. Goldstein'ın da belirttiği gibi, ICD kodlarının kullanımı, salgınların tespiti ve hasta problem listelerinin geliştirilmesi ve netleştirilmesinde önemli rol üstlenmektedir. Ayrıca ICD kod kullanımı faturalama hizmetlerinde de standardizasyon sağladığından sağlık kurumlarına ciddi fayda sağlamaktadır [1].

Doğru kodların raporlara atanması işi hem emek yoğun ve hem de ciddi konsantrasyon ve geniş bir bilgi birikimi gerektiren karmaşık bir süreçtir. Kodları atayan kişi binlerce kod içinden sadece az sayıda ilgili kodu raporla ilişkilendirmelidir. Karmaşıklığı nedeniyle hataya açık bir işlemdir ve uzman kişiler tarafından gerçekleştirilmelidir. ICD-10 kodlarının klinik raporlara otomatik atanması işi kodlayıcılar için süreci kolaylaştırması, hata oranını azaltması açısından memnuniyet verici olacaktır. Bu problem tanımı çerçevesinde bu çalışmanın ICD-10 kodlarını atayan uzmanlara yardımcı olmak amacıyla, verilen bir tıbbi kayıtta yer alan serbest formda yazılmış tıbbi metin sahaları kullanılarak atanacak ICD-10 kodlarını önerecek akıllı bir sistem geliştirmektir. Bu amaç doğrultusunda, çalışmada bilgi gerigetirme (information retrieval) tekniklerine dayalı ve farklı parametrelerle başarımları yönetilebilen yeni bir metin sınıflandırma metodu önerilmiştir.

2. Gereç ve Yöntem

ICD-10 kodları, hastalıkların tüm genel epidemiyolojik, birçok sağlık yönetimi amaçları ve klinik kullanımı için oluşturulmuş uluslararası tanısal sınıflandırma standardıdır. Hasta takibi, hasta takip ve arşivinin tutulması ve kaynak yönetimi gibi idareye yönelik amaçların yanı sıra hastalıklarla ilgili istatistiksel veriye dayanması ve uluslararası olması açısından ülkeler arasında da sağlıkla ilgili karşılaştırmalar yapılmasına olanak sağlamaktadır. ICD-10 sınıflandırmasını teknik bir sınıflandırma yönteminden ziyade istatistiksel bir yöntem olarak görmek gerekmektedir. ICD-10 kod standardı ölüm sertifikaları ve sağlık kayıtları da dâhil olmak üzere sağlık ve hayati kayıtların birçok türünü, kaydedilen hastalıkları ve diğer sağlık problemlerini sınıflandırmayı amaçlar; bu amaç dâhilinde kayıtların daha güvenilir ve doğru olmasını sağlar. ICD-10 kod yapısı; XXX.YY, 1. Seviye XXX (A00-Z99), 2.Seviye YY (01-09); şeklindedir. ICD-10 kod başlıkları ve bir gruba ait alt başlıklar **Hata! Başvuru kaynağı bulunamadı.**'de verilmiştir.

Bu çalışmada Türkçe tıbbi metinlerin işlenerek ICD-10 hastalık kodlarının raporlara atanmasına yardımcı olacak akıllı bir sistem önerilmiştir. Geliştirilen sistem, metin sınıflandırma için yeni bir yöntem de içermektedir. Metin sınıflama işlemi kabaca bir dokümanın belirli bir sınıfa veya kategoriye atanması işlemidir. Önerilen metin sınıflandırma yönteminin ayrıntılarını açıklamaya geçmeden önce literatürde metin sınıflandırma için önerilen belli başlı yaklaşım ve algoritmaları kısaca özetlemek istiyoruz.

Çoklu kategori için metin sınıflandırma alanında en yaygın yöntemlerden birisi olan Rocchio algoritmasını Miao ile Kamel geliştirmişler ve farklı alanlarda uygulamışlardır (Miao & Kamel, 2011). Zhang, Shuigeng Zhou, ve Aoying Zhou (Zhang, Zhou, & Zhou, 2004) sıralı sınıflandırıcılar üretmek ve bu sınıflayıcıları filtreleme ve karar verme aşamasında kullanmak için Arttırımlı (Boosting), KNN ve

Rocchio sınıflandırıcıları birleştirmiş ve Çince haberlerden oluşan veri setlerine uygulamışlardır.

Tıbbi alanda metin sınıflandırma ise bazı noktalarda farklılık gösterir. Tıbbi metin sınıflandırma süreci dokümanların biyomedikal çeşitliliğe veya disipline göre etiketlenmesinden oluşur. Tıbbi raporlara ICD, MeSH, SNOMed, GO gibi kategorilerin kodlarının atanması, tıbbi metin sınıflandırma görevlerinin içinde yer alsa da, kodların yapısının farklılığı nedeni ile tıbbi metin sınıflandırmadan ayrı tutulmalıdır (Humphrey, Névéol, Gobeil, Ruch, Darmoni, & Browne, 2009). Bir tıbbi metin sınıflandırma uygulaması olarak Humphrey ve diğerleri kural tabanlı bir sistem ile istatistiksel bir sistemi MEDLINE dokümanlarını biyomedikal çeşitliliğe göre otomatik sınıflandırmaya çalışmışlardır. İki yöntemin performansı birçok yönden karşılaştırabilir olmakla birlikte, iki yöntemin birleştirilmesi ile daha iyi sonuçlar elde edilebildiği gösterilmiştir.

Tablo 1 - Hastalıklar ve Sakatlıklar Tabular Endeksi (Seviye-1) & (Seviye -2).

```
1. Bölüm Enfeksiyon ve Paraziter Hastalıklar (A00 - B99)
2. Bölüm Neoplazmlar (C00 ile D48)
3. Bölüm Kan ve Kan Yapıcı Organ Hastalıkları ve Bağışıklık Sistemini İçeren
Hastalıklar (D50-D89)
   1.  D50-D53: Besinsel anemiler
       1.1  D50 Demir eksikliği anemisi
            D50.0  Kronik kan kaybına bağlı demir eksikliği anemisi
            D50.1  Sideropenik disfaji
            D50.8  Diğer demir eksikliği anemileri
            D50.9  Belirlenmemiş demir eksikliği anemisi
       1.2  D51 B12 vitamin eksikliği anemisi
       1.3  D52 Folat eksikliği anemisi
       1.4  D53 Diğer besinsel anemiler
   2.  D55-D59: Hemolitik anemiler
   3.  D60-D64: Aplastik ve diğer anemiler
   4.  D65-D69: Koagülasyon bozuklukları, purpura ve diğer hemorajik durumlar
   5.  D70-D77: Kan ve kan oluşturan organların diğer hastalıkları
   6.  D80-D89: Bağışıklık sistemini içeren belirli bozukluklar
4. Bölüm Endokrin, Nutrisyonel ve Metabolik Hastalıklar (E00-E90)
5. Bölüm Akıl ve Davranış Bozuklukları (F00-F99)
6. Bölüm Sinir Sistemi Hastalıkları (G00-G99)
7. Bölüm Göz ve Gözle Bağlantılı Doku Hastalıkları (H00-H49)
8. Bölüm Kulak ve Mastoid Oluşum Hastalıkları (H60-H95)
9. Bölüm Dolaşım Sistemi Hastalıkları (I00-I99)
10. Bölüm Solunum Sistemi Hastalıkları (J00-J99)
11. Bölüm Sindirim Sistemi Hastalıkları (K00-K93)
12. Bölüm Cilt ve Cilt altı Dokusu Hastalıkları(L00-L99)
13. Bölüm Kas-İskelet ve Bağ Dokusu Hastalıkları (M00-M99)
14. Bölüm Ürogenital Sistem Hastalıkları (N00-N99)
15. Bölüm Gebelik, Doğum ve Lohusalık Dönemi Hastalıkları (O00-O99)
16. Bölüm Perinatal Dönemden Kaynaklanan Hastalıklar (P00-P96)
17. Bölüm Konjenital Malformasyon, Deformasyon ve Kromozom Anomalileri (Q00-Q99)
18. Bölüm Semptomlar ve Anormal Klinik ve Laboratuar Bulguları (R00-R99)
19. Bölüm Yaralanma, Zehirlenme ve Dış Nedenlere Bağlı Diğer Durumlar (S00-T98)
20. Bölüm Hastalık ve Ölümün Dış Nedenleri (V01-Y98)
21. Bölüm Sağlık Durumu ve Sağlık Hiz. Yararlanmayı Etkileyen Faktörler (Z00-Z99)
```

Yi & Beheshti ise [5] Hidden Markov Model (HMM) tekniğini tıbbi dokümanları 23 MeSH Kategori C'nin alt kategorilerine atamak için kullanmışlardır. Bu çalışmanın sonuçlarının literatürde yer alan diğer yöntemler ile karşılaştırılabilir olduğunu belirtmişlerdir. Moore ve Berman [6] SNOMED kodlarını otomatik atayacak bir sistem geliştirmek üzerine çalışmışlar ve otomatik kodlamanın birçok açıdan elle kodlamaya

üstünlük sağladığı üzerinde durmuşlardır. Yaklaşımları metinler içindeki bariyer ve ana kelimeleri çözümlemeye dayanmaktadır ve çözümleme sonucunda elde edilecek ana kelimeler SNOMED kodlarına işaret eder. Popa ve arkadaşları [7] çoklu-etiket ve çoklu kategori için tıbbi doküman sınıflandırma probleminin çözümünde bir denetleyici öğrenme yöntemi olan matris regresyon metodunu önermiştir. Öğrenme aşamasında sistem kelime-kategori matrisi oluşturmakta, yeni dokümanların ICD kodlarının sınıflandırılmasında ise puanlama için bu matrisi kullanmaktadır. Bu yöntem ile performans ölçütü olarak f-ölçeğini 0.34 hesaplamışlar ve bu yöntemin SVM, k-NN ve Centroid yöntemlerinden daha iyi sonuç verdiğini belirtmişlerdir. Crammer ve arkadaşları, ICD-9-CM kodlarının serbest formda yazılmış radyoloji raporlarına atanabilmesi için bir sistem geliştirmişlerdir. Öğrenme seti 978, test seti ise 976 etiketsiz tıbbi dokümandan oluşmaktadır. Radyoloji raporları klinik geçmiş ve izlenim olmak üzere iki alan içermektedir. Bu raporlar 45 ayrı ICD kodu ve oluşturduğu 94 yapılandırmayı içermektedir. Bu çalışmalarında üç farklı kodlama sistemini doğruluk performans değerini yükseltmek amacıyla tek bir öğrenme sisteminde birleştirmişlerdir. Öncelikli olarak otomatik kodlama prensibini ve kural bazlı atamayı kullanmışlar, eğer sistemler sonuç üretmez ise öğrenme yöntemini uygulamışlardır. f-performans değerini %86.5 elde etmişlerdir, aynı veri seti için elde edilmiş performans ise % 89.08 olarak rapor edilmiştir. [8].

Benzer olarak Goldstein ve arkadaşları CMC için ICD-9-CM kodlarını radyoloji raporlarına atamak için üç farklı yöntem denemişlerdir. Öncelikli olarak ayrık kelimelerin benzerliği ilkesine dayanarak Lucene bilgi gerigetirim aracını kullanarak atama yapmışlardır. İkinci olarak BoostTexter isimli *n*-gram (ardışık kelime dizileri) ve *s*-gram (ardışık olmayan kelime dizileri) dayanan artırımlı bir algoritma yöntemini uygulamışlardır. Son olarak ise olumsuz ifadeleri, eş anlamlı ifadeleri ve kesin olmayan verileri ayırt edebilecek kural tabanlı bir sistem geliştirmişler ve basit kural tabanlı bir sistemin f-performans değeri %88.55 ile diğerlerinden daha iyi sonuç verdiğini ortaya koymuşlardır [1]. Chen ve diğerleri [9] bağımlılık ayrıştırma prensibine dayalı anlamsal analitik tekniği ile radyoloji raporlarına otomatik ICD-9-CM kodlarının atanmasını amaçlamıştır. Yaklaşımlarını 2007 International Natural Language Processing Challenge veri seti üzerinde test etmişlerdir. Bu çalışmanın ana fikri bağımlılık ayrıştırma, ayrıştırma ağacı eşleme ve eğitim ve sınama veritabanı arasındaki anlamsal eşleme puanı hesaplamayı içeren derin düzeyde semantik analizi yaklaşımıdır. Macro-average F1 performans değerini 0.60 olarak yakalamışlardır, Goldstein, Arzumtsyan ve Uzuner [1] bu değeri aynı veri setinde 0.73 olarak belirtmişlerdir. Buna bağlı olarak Chen ve arkadaşları kural tabanlı sistemlerin daha iyi sonuç vermesine rağmen bilgi edinme darboğazı nedeniyle uygulamada zayıf yanları olduğunu savunmaktadır.

Boytcheva, Bulgarca için, endokrin ve metabolik hastalıkların serbest metin çıkış raporlarından ICD-10 kodlarının otomatik atanmasında çoklu sınıf için SVM tekniğini kullanmıştır. Problem çoklu ikili sınıflandırma problemine indirgenmiş ve maks-min oylama stratejisi ile gerçeklenmiştir, çalışma sonucunda f-performans değeri %84,5 olarak bildirilmiştir [10]. Son olarak, Foudeh ve Salim [11] bir ICD sınıfını karakterize edecek anlamlı kelimeleri bulmayı amaçlayan kelime ağırlıklandırma modelini önermişlerdir.

Yöntem

Metin sınıflandırma, dokümanda yer alan metinlerin kategorilerinin belirlenmesi için önceden belirli olan uygun etiketlere atanmasını amaçlar. Tıbbi metin sınıflandırma açısından, serbest metinler tıbbi alanın temel bilgi kaynağıdır. Bu tür veriler karar verme sistemlerinde yapısal ve kodlanmış veri kadar yararlı değildir, beraberinde bu verinin tekrar kullanılabilmesi tekrar çaba harcanmasını gerektirir. Bu bağlamda bu çalışmada amaç, tıbbi raporların kullanılarak birinci seviye ICD-10 kodlarına göre sınıflandırılarak serbest metin verisinin kullanılabilir olmasını sağlamaktır.

Tıbbi metin sınıflandırma diğer metin sınıflandırma süreçleri arasında daha karmaşık bir yapıya sahiptir bunun nedeni tıbbi metin sınıflandırmanın çok kategorili, birçok kategori yapısına uyan bir sınıflandırma işlemi olmasıdır. Bunun yanı sıra metinlere kod atama işleminde Crammer ve diğerlerinin [8] de değindiği üzere; bir teşhis kodu ancak ilgili teşhise ulaşıldığında atanmalı, bir teşhis kodu teşhis net olmadığı durumda asla atanmamalı, bir belirti kodu teşhis kesinleştiğinde atanmamalı ve her zaman en özellikli kod tercih edilmelidir. Başka bir anlamda uygun görünen kodlar özel durumlarda dâhil edilmemelidir. Bu zorlukların yanı sıra tıbbi rapor sınıflandırma süreçlerinin eşanlamlı ve olumsuzluk belirten ifadelerden de etkilendiği unutulmamalıdır.

Bu çalışmada geliştirilen sistem, kodlama yapacak uzmana, verilen bir tıbbi kayıtta yer alan serbest formda yazılmış metin sahaları işlenerek, atanması muhtemel ICD kodları önermektedir. Sistem temel olarak bilgi gerigetirim sistemlerine dayalı bir yaklaşıma sahiptir. Bu yaklaşımın en önemli bileşeni bilgi gerigetirim sistemidir ve bu çalışmada Glaskow Üniversitesi tarafından geliştirilen ve açık kaynak kodlu olarak toplumun kullanımına sunulmuş olan Terrier bilgi gerigetirim sistemi [**12**] kullanılmıştır. Terrier bilgi gerigetirim sistemi, simge normalleştirme vektör uzay matrisi oluşturma, endeksleme ve geri getirim işlemlerini gerçekleme gibi özelliklere sahip bir araçtır.

Bilgi gerigetirim sistemlerinin temelini oluşturan vektör uzayı modelinde, tüm veri seti Terim Doküman Matrisi (TDM) adı verilern $m \times n$'lik bir matris ile temsil edilir. TDM matrisinde, sütunlar doküman setindeki bir dokümanı, satırlar ise bir dokümandaki farklı bir kelimeyi temsil eder. TDM matrisinin verilen bir elemanını, a_{ij}, j'ninci dokümanın i'ninci terimini ifade eder. Bu değer, kaba olarak bir kelimenin ilgili dokümanda kaç kere geçtiğini belirtir. Bir dokümanda ilgili kelime geçmiyor ise bu değer 0 olacaktır. Her bir matris elemanının değeri farklı yöntemlerle ağırlıklandırılabilir. Kelime kümesi $T=\{t_1,t_2,\ldots,t_m\}$ ve doküman kümesi de $D=\{d_1,d_2,\ldots,d_n\}$ ile temsil edilecek olur ise a_{ij}; t_i kelimesinin d_j dokümanındaki frekansı veya farklı parametrelere göre hesaplanmış ağırlık değeridir. Belirtilen tanım dahilinde Terrier bir sorgu vektörü $d_q=\{a_{0q},a_{1q},\ldots,a_{mq}\}$ ile bir doküman vektörü arasındaki $d_j=\{a_{0j},a_{1j},\ldots,a_{mj}\}$ benzerlik değerini $\sum_{i=0}^{m} a_{ij} \times a_{iq}$ formülü ile hesaplar. a_{iq} ve a_{ij} değerleri kelimelerin frekans değerleri değil farklı terim ağırlıklandırma yöntemlerine göre hesaplanmış değerlerdir. Bilgi Gerigetirme tekniklerine dayalı önerilen sınıflandırma yönteminin adımları aşağıda gösterilmiştir.

1. Eğitim dokümanlarını işaretleyerek sisteme yükle
   Eğitim Seti $ES = \{d_0, \ldots, d_l\}$ l = eğitim setindeki dokuman sayısı
   $ES \in DE$ (Doküman endeksi)
2. Test dokümanlarının her bir doküman ile benzerliğini hesapla
   Test Seti $TS = \{q_0, \ldots, q_p\}$ p = test setindeki doküman sayısı

q sorgusunun j dokümanı ile benzerliği $B_{qj} = \sum_{i=0}^{m} a_{ij} \times a_{iq}$

3. İlgili dokümanların listelendiği azalan sırada sıralı sonuç kümesinin ilk 10 dokümanını al.
   Sonuç Kümesi $RS = \{d_0, \ldots, d_{1000}\}$ Sonuç kümesi sıralı dokümanların en yüksek benzerliğe sahip 1000 tanesini içerir. Sonuç alt kümesi $RSS = \{d_0, \ldots, d_r\}$ $r = 10$.
4. Sorgulara ICD kodları atamak için *RSS*'yi ağırlıklandır
   Bir dokümanının ICD kod kümesi, $IK = \{ICD_0, \ldots, ICD_z\}$
   $W_{ICDZ} \leftarrow$ Ağırlıklandır(ICD_z)
5. Aday sınıfları ağırlıklarına göre sırala
   Sıralı ICD Kod Kümesi (*IKS*) ← AzalanSıradaSırala(W_{ICD})
6. Sonuç sınıf kümesi (*SSK*) ← SonuçSınıfKümesiSeç(*IKS*)

Bilgi gerigetirme tekniklerine dayalı metin sınıflandırma yöntemin ICD sınıf ataması yapan 4. adımında yer alan Ağırlıklandır(ICD_z) fonksiyonu iki farklı tekniğe göre ağırlık hesaplamaktadır. Birinci teknikte amacımız, dokümanın sıra endeksine göre ağırlık değeri hesaplamaktır: Eğer bir doküman ilk sırada geldi ise ağırlık faktörü 10, eğer bir doküman 10. Sırada geldi ise dokümanın ICD kodlarının ağırlık faktörü 1'dir. r. sırada gelen dokümanın ağırlık faktörü $wd_r = 11 - r$ formülü ile belirlenir. Bir ICD sınıfı için ağırlık değeri ilk 10 dokümanda yer alan tüm ICD'ler için wd_r değerlerinin toplanması ile elde edilir $\forall\ W_{ICD} = \sum wdr$. Uygulanan ikinci teknik ise her ICD kodunu doküman benzerlik verisine göre ağırlıklandırmayı esas alır. İlk 10 dokümanda yer alan tüm ICD'ler için doküman sorgu benzerlik değerleri ($B_{qr)}$ toplanır $\forall\ W_{ICD} = \sum Bqr$. Çalışmamızın ileriki aşamalarında amaçlanan ICD sınıf atamalarında uygulanabilecek farklı ağırlıklandırma tekniklerini geliştirmektir.

Yöntemin son adımında yer alan SonuçSınıfKümesiSeç(*IKS*) yordamı her bir sorgu sınıf dokümanı için ağırlıklarına göre sıralanmış ICD sınıfları içerisinden aday sınıfların ilk k tanesini alarak sorgu dokümanın sınıflarını atar, $SSK = \{ICD_0, \ldots, ICD_K\}$ $k=\{1,\ldots,10\}$. Bu aşamada uyguladığımız diğer yol ise belirlenen bir eşik (t) seviyesinin üzerinde ağırlığa sahip olan ICD sınıflarının sorgu dokümanına atanmasıdır, $SSK = \{W_{ICD} > (t)\}$ $t = \{0,100,500,100\}$

3. Bulgular

Veri Setinin Hazırlanması

Bu çalışmada, Ankara Üniversitesi Tıp Fakültesi Hastanesi'nin 2010 yılının 11 ayında yatan hastalara ait tıbbi kayıtlar kullanılmıştır. Kullanılan veriler 11 ayrı Excel dosyası formatında bilgi sisteminde alınmıştır. Her bir dosyada dosya numarası, hasta adı soyadı, rapor tarihi, yatış-çıkış tarihleri gibi sabit veriler ile ICD hastalık kodları yer almaktadır. Ayrıca, her bir kayıtta "anamnez, fizik inceleme, klinik seyir, ameliyat raporu" gibi serbest formda yazılmış sahalar yer almaktadır. Dosya no sahası, yeniden düzenlenerek hastanın kimliğini ifşa etmeyecek şekilde bir sıra numarası haline dönüştürülmüştür. *Ad-soyad* özniteliği XML dosyalarına eklenmemiş, ayrıca isim bilgisi içeren bazı sahalardan özel isim ve benzeri kişisel bilgiler metinlerden temizlenerek veriler anonimize edilmiştir.

Bu çalışmada, ilk aşamada toplam 57835 hasta kaydı seçilmiştir. Ardından, anamnez, fizik inceleme ve klinik seyir gibi sahalardaki verilerin metin uzunlukları

10’dan küçük olan kayıtlar ile geçerli bir ICD-10 kodu bulunmayan kayıtlar da sistemden çıkarılmıştır. Tüm bu seçme ve düzenleme işlemleri ardından 57835 kayıttan 52100 adedi geçerli kayıt olarak belirlenmiş ve her bir kayıt için

Şekil 1’de gösterilen formatta XML dosyaları oluşturulmuştur. Hazırlanan veri kümesinden 145 tanesi sınama kümesi, 51955 tanesi ise eğitim kümesi olarak seçilmiştir. Sınama kümesinde yer alan 145 kayıtta toplamda 486 olmak üzere 133 farklı ICD kodu bulunmaktadır.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<report>
<_ReportNo>10010</_ReportNo>
<Anamnez>şikayeti:göğüs ağrısı  hikayesi:3 aydır şikayetleri
var.2 kat merdiven çıkmakla göğüs ağrısı ve nefes darlığı var
efor kapasitesi progresif olarak azalmış.baskı tarzında olan
ağrısı 1-2 dakikada geçiyor.hasta ileri tetkik vetedavi için
kliniğe yatırıldı.</Anamnez>
<FizikInceleme>KVS:s1 normal s2 normal sistolik apikal üfürüm
1/6.ek ses yok. solunum:normal  GİS:doğal /op skarı.
ekstremite:nabızlar açık.ptö-/-
</FizikInceleme>

<KlinikSeyir>hasta kliniğe kabul edildi.medikal tedavisi
düzenlendi.rutin tetkikleri istendi.hastaya 25/02/2010 da KAG
yapıldı.LAD proksimalde % 30 darlık izlendi,LAD ortada % 80
darlık izlendi.D1 ortada % 50 darlık izlendi.Cx ostiumda % 30
darlık izlendi,ortada % 70 darlık izlendi.Majör obtus ortada %
90 tübüler darlık izlendi.Majör obtus proksimalinden çıkan yan
dal proksimalinde % 95 tübüler darlık izlendi.RCA ektazik
izlendi.CABG kararı alınan hasta KVCye ayaktan yatışı yapılmak
üzere taburcu edildi.</KlinikSeyir>
<Teshis>
 I25 1 ATEROSKLEROTİK KALP HASTALIĞI(KORONER (ARTER)
HASTALIK)
 I10 * ESANSİYEL (PRİMER) HİPERTANSİYON
 E78 2 HİPERLİPİDEMİ KARMA
 E11 * İNSÜLİN-BAĞIMLI OLMAYAN DİYABETES MELLİTÜS
</Teshis>
<Rapor>ilaç raporu verildi.</Rapor>
<Sonuc>taburcu.</Sonuc>
.
.
</report>
```

Şekil 1 – XML formatında oluşturulan tıbbi kayıt dosyası örneği.

Veri Setine ait ICD kod istatistikleri

Bu bölümde koleksiyona ait verilerden elde edilen basit istatistiksel değerlere yer verilmiştir. İlk istatistik *Tablo 2*’de yer alan her bir dokümanın kaç ICD sınıfı içerdiğini gösterir.

Tablo 2 - Her bir dokümanın içerdiği ICD kod sayısına göre tıbbi kayıt sayıları.

Vaka Başına ICD kod sayısı	Vaka Sayısı
1	25818
2	9656
3	5450
4	3810
5	2873
6	2346
7	1494
8	549
9	97
10	5
11	2

Bir başka istatistik ise, aynı tıbbi dokümanda yer alan ICD sınıflarının ilişkisidir. En sık oranda geçen 30 adet ikili ICD sınıfının bilgisi *Tablo 3*'de verilmiştir. Bu istatistik, tıbbi alanda bilgi çıkarsama ve ICD'ler arası bağ oluşturma için bir başlangıç adımı olarak kabul edilebilir.

Tablo 3 – ICD kodlarının aynı dokümanda birlikte görünme sıklıkları.

ICD1	ICD2	Vaka Sayısı	ICD1	ICD2	Vaka Sayısı
I25	I10	3324	I48	I10	676
E78	I10	2665	I25	E11	635
I25	E78	2290	I50	J44	625
I10	E11	1150	E78	E11	605
I10	K29	1119	E11	E78	605
I10	J44	1057	N18	I10	580
J44	I10	1057	Z95	I25	549
I50	I10	984	I25	I20	543
I10	E10	896	I10	Z95	532
I25	K29	860	E10	E78	507
I50	I25	800	I25	I48	496
O82	Z33	777	I50	I48	481
J44	J96	769	I25	N18	480
I25	J44	758	I25	E10	465
E78	K29	686	I20	I10	437

Performans Ölçekleri

Bu bölümde sınıflandırma yöntemlerine ait performans hesaplama ölçütlerinin temel bir tanımı verilmiştir: Doğru Pozitif değerler (true positive - TP) bir sınıflandırma sistemi için doğru tahmin sayısını Doğru Negatif değerler (true negative - TN) ise doğru ret Sayısını temsil eder sistemin hata parametreleri ise yanlış atamalar için (1. Tip Hata) Yanlış Positif değerler (false positive - FP), kayıp veriler (2.Tip Hata) için ise Yanlış Negatif değerler (false negative - FN) ile temsil edilir. Bu parametrelere bağlı olarak sistemin duyarlılığı (sensitivity : TP/(TP+FN)), doğruluğu

(accuracy : ((TP+TN) /(TP+TN+FP+FN)) ve özgüllüğü (specificity : TN / (FP+TN)) hesaplanabilir.

Performans Değerleri

Uygulanan yönteme bağlı olarak Tablo 4 ve Tablo 5 farklı test parametrelerine göre test sonuçlarını göstermektedir. Yer alan performans değerleri: Başarı ile getirilen sorgu sayısı (eğer bir sorguya ait en az bir tane ICD kodu başarı ile atanmış ise sorgu başarılı sayılmıştır), başarı ile getirilmiş ICD sayısı, toplam döndürülen ICD Sayısı ve bu değerlere bağlı olarak hesaplanmış, Doğru atanmış ICD sayısı oranı, duyarlılık, doğruluk ve özgüllük parametreleridir.

Tablo 4–Satır endeksine göre ağırlıklandırılmış test sonuçları.

k / t	Başarılı Sorgu#	Başarılı ICD#	Toplam Getirilen	Sorgu bazında doğru atanmış ICD Oranı (%)	Sınıf bazında duyarlılık	Sınıf bazında özgüllük	Sınıf bazında doğruluk	Sorgu bazında duyarlılık	Sorgu bazında özgüllük	Sorgu bazında doğruluk
1	111	111	145	36	0,229	0,989	0,922	0,363	0,998	0,977
2	125	173	290	50	0,356	0,973	0,927	0,497	0,992	0,976
3	132	217	428	57	0,447	0,955	0,925	0,571	0,986	0,974
4	134	248	564	62	0,509	0,936	0,92	0,620	0,983	0,970
5	137	272	694	66	0,559	0,925	0,916	0,656	0,978	0,966
6	139	286	818	68	0,588	0,914	0,908	0,676	0,971	0,961
7	140	296	930	69	0,61	0,903	0,902	0,690	0,966	0,956
8	140	304	1037	70	0,626	0,896	0,896	0,704	0,961	0,952
9	140	310	1144	71	0,638	0,886	0,89	0,713	0,956	0,946
10	140	319	1245	73	0,656	0,88	0,886	0,725	0,950	0,942

Tablo 5 - Doküman benzerliğine göre ağırlıklandırılmış göre test sonuçları.

k / t	Başarılı Sorgu#	Başarılı ICD#	Toplam Getirilen	Sorgu bazında doğru atanmış ICD Oranı (%)	Sınıf bazında duyarlılık	Sınıf bazında özgüllük	Sınıf bazında doğruluk	Sorgu bazında duyarlılık	Sorgu bazında özgüllük	Sorgu bazında doğruluk
1	112	112	145	36	0,231	0,986	0,921	0,364	0,998	0,978
2	126	176	290	50	0,362	0,972	0,928	0,503	0,992	0,977
3	132	219	428	57	0,451	0,95	0,922	0,575	0,986	0,974
4	136	252	564	63	0,519	0,935	0,921	0,627	0,983	0,971
5	137	273	694	65	0,562	0,923	0,914	0,655	0,978	0,966
6	139	286	818	68	0,588	0,912	0,908	0,677	0,971	0,961
7	139	292	930	68	0,601	0,897	0,896	0,684	0,965	0,956
8	140	306	1037	71	0,629	0,889	0,892	0,708	0,961	0,952
9	140	315	1144	72	0,648	0,882	0,887	0,719	0,956	0,947
10	141	321	1245	73	0,661	0,874	0,88	0,725	0,951	0,942
t >0	142	372	2430	79	0,765	0,816	0,837	0,794	0,890	0,887
t>100	134	331	1476	71	0,681	0,874	0,878	0,709	0,938	0,932

t >500	85	157	234	36	0,323	0,977	0,931	0,362	0,995	0,977
t >1000	42	55	65	17	0,113	0,997	0,916	0,173	0,999	0,976

Ek olarak, Tablo 6 test seti içerisinde yer alan en sıklıkla görülen 40 ICD sınıfı için performans değerlerini gösterir.

Tablo *7* sorgu bazlı hesaplanmış performans değerlerini içermektedir. İlk 20 sorgu için performans sonuçları Tablo 7'de verilmiştir .

Tablo 6 – Veri setinde en sıklıkla görülen 40 ICD için sınıflandırma performans değerleri (Sonuçlar doküman benzerliğine göre ağırlıklandırma t > 0 testini temel alır.)

ICD kodu	Sorgu #	İlgili getirilen	Getirilen	Sıklık	TP	FN	FP	TN	Duyarlılık	Özgüllük
I10	46	43	94	7993	43	3	51	48	0,93	0,48
I25	28	28	73	4972	28	0	45	72	1	0,62
E78	23	22	63	3441	22	1	41	81	0,96	0,66
J44	29	24	57	2470	24	5	33	83	0,83	0,72
K29	18	15	46	2101	15	3	31	96	0,83	0,76
Z33	10	10	20	1861	10	0	10	125	1	0,93
I50	15	14	42	1802	14	1	28	102	0,93	0,78
E11	10	7	62	1696	7	3	55	80	0,7	0,59
N18	10	8	38	1643	8	2	30	105	0,8	0,78
J96	12	12	35	1570	12	0	23	110	1	0,83
E10	12	9	44	1566	9	3	35	98	0,75	0,74
J18	13	13	44	1417	13	0	31	101	1	0,77
O82	10	10	17	1248	10	0	7	128	1	0,95
I48	10	9	43	1185	9	1	34	101	0,9	0,75
C34	10	10	18	1001	10	0	8	127	1	0,94
Z95	10	6	21	912	6	4	15	120	0,6	0,89
C18	10	10	16	907	10	0	6	129	1	0,96
F32	10	8	21	904	8	2	13	122	0,8	0,9
K80	10	9	14	859	9	1	5	130	0,9	0,96
H25	10	8	8	816	8	2	0	135	0,8	1
K21	2	1	24	794	1	1	23	120	0,5	0,84
R10	2	0	8	733	0	2	8	135	0	0,94
N40	4	4	22	720	4	0	18	123	1	0,87
I67	2	2	13	714	2	0	11	132	1	0,92
N20	2	1	3	698	1	1	2	141	0,5	0,99
I20	1	1	14	677	1	0	13	131	1	0,91
N39	1	0	11	663	0	1	11	133	0	0,92
D64	2	1	16	647	1	1	15	128	0,5	0,9
Z94	2	2	7	644	2	0	5	138	1	0,97
C67	1	1	3	636	1	0	2	142	1	0,99
E03	6	2	25	614	2	4	23	116	0,33	0,83
E14	2	0	15	586	0	2	15	128	0	0,9
N17	4	1	16	581	1	3	15	126	0,25	0,89
M81	4	3	16	553	3	1	13	128	0,75	0,91
D51	2	1	16	541	1	1	15	128	0,5	0,9

J45	2	0	19	531	0	2	19	124	0	0,87
B18	1	1	16	527	1	0	15	129	1	0,9
K74	1	1	13	523	1	0	12	132	1	0,92
Y45	4	2	14	478	2	2	12	129	0,5	0,91

Tablo 7 - Sorgu bazında performans değerleri - Sonuçlar doküman benzerliğine göre ağırlıklandırma $t > 0$ testini temel alır

Sorgu No	TP	FN	FP	Duyarlılık	Özgüllük	Doğruluk
1	2	0	4	1	0,97	0,97
2	4	1	20	0,8	0,84	0,84
3	1	3	14	0,25	0,89	0,87
4	2	0	8	1	0,94	0,94
5	0	2	17	0	0,87	0,86
6	1	1	10	0,5	0,92	0,92
7	3	0	22	1	0,83	0,83
8	2	0	18	1	0,86	0,86
9	5	0	14	1	0,89	0,89
10	4	0	10	1	0,92	0,92
11	4	2	17	0,67	0,87	0,86
12	1	0	6	1	0,95	0,95
13	5	0	9	1	0,93	0,93
14	2	0	20	1	0,85	0,85
15	4	4	23	0,5	0,82	0,8
16	4	0	12	1	0,91	0,91
17	4	0	6	1	0,95	0,95
18	6	1	19	0,86	0,85	0,85
19	2	0	12	1	0,91	0,91
20	4	0	21	1	0,84	0,84

4. Tartışma ve Sonuç

Bu çalışmada bilgi gerigetirme tabanlı sınıflandırma yöntemi kullanılarak tıbbi kayıtlara ICD-10 kodu atanmasına yardımcı bir akıllı sistem önerilmiştir. Sistem, Ankara Üniversitesi Tıp Fakültesi Hastanesi yatan hasta kayıtlarından derlenen 51955 kayıt ile eğitilmiş ve 145 sorgu ile test edilmiştir. Yapılan deneylerden elde edilen sonuçları doğrultusunda en iyi duyarlılık değeri 0,765 olarak doküman benzerliğine göre ağırlıklandırma yönteminde $t > 0$ alındığı durumda elde edilmiştir.

Sınıf bazlı sorgu sonuçlarında dalgalanmalar yer alsa da bunun nedenleri arasında sınıfların eğitim seti içerisindeki örnek sayısına veya bir sınıfın sistem tarafından başka yakın sınıflar ile karıştırılabilmesine bağlı olduğu düşünülebilir. Farklı veri setleri üzerinde çapraz doğrulama yapılarak daha anlamlı sonuçlar elde edilebilir. Oluşturulan testler ve uygulanan parametreler yeni geliştirilen bilgi gerigetirme tekniklerine dayalı sınıflandırma algoritmasına temel oluşturmak amacıyla düzenlenmiştir. Sorgu bazlı performans değerlerine göre, çoğu sorgu açısından duyarlılık değerleri kabul edilebilir ölçektedir. k değerinin artmasına bağlı olarak duyarlılık performansı büyük oranda artmaktadır, bu artış ile birlikte özgüllük ve doğruluk değerlerinde azalma gözlense de azalma aynı hızla olmamakta ve kabul edilebilir değerlerin altına inmemektedir. Bu

aşamada yöntemin geliştirilmesi için bundan sonraki amaç, var olan parametrelerin optimizasyonu ve yeni parametrelerin sisteme dâhil edilmesi olarak planlanmıştır.

5. Kaynakça

[1] I. Goldstein, A. Arzumtsyan and Ö. Uzuner, "Three Approaches to Automatic Assignment of ICD-9-CM Codes to Radiology Reports," *AMIA Annu Symp Proc*, p. 279–283, 2007.

[2] Y.-Q. Miao and M. Kamel, "Pairwise optimized Rocchio algorithm for text categorization," *Pattern Recognition Letters 32*, pp. 375-382, 2011.

[3] Z. Zhang, S. Zhou and A. Zhou, "Sequential Classifiers Combination for Text Categorization: An Experimental Study," *Advances in Web-Age Information Management Lecture Notes in Computer Science, Volume 3129/2004*, pp. 509-518, 2004.

[4] S. M. Humphrey, A. Névéol, J. Gobeil, P. Ruch, S. J. Darmoni and A. Browne, "Comparing a Rule Based vs. Statistical System for Automatic Categorization of MEDLINE® Documents According to Biomedical Specialty," *Journal of the American Society for Information Science and Technology*, pp. 2530-2539, 2009.

[5] K. Yi and J. Beheshti, "A hidden Markov model-based text classification of medical documents," *Journal of Information Science*, pp. 67-81, 2009.

[6] W. G. Moore and J. J. Berman, "Automatic SNOMED coding," in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 1994.

[7] I. Sandu Popa, K. Zeitouni, G. Gardarin, D. Nakache and E. Metais, "Text Categorization for Multi-label Documents and Many Categories," *Proceeding CBMS '07 Proceedings of the Twentieth IEEE International Symposium on Computer-Based Medical Systems* , 2007.

[8] K. Crammer, M. Dredze, K. Ganchev, P. P. Talukdar and S. Carroll, "Automatic code assignment to medical text," *Proceeding BioNLP '07 Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, 2007.

[9] P. Chen, A. Barrera and C. Rhodes, "Semantic analysis of free text and its application on automatically assigning ICD-9-CM codes to patient records," *Cognitive Informatics (ICCI), 2010 9th IEEE International Conference*, pp. 68 - 74, 2010.

[10] S. Boytcheva, "Automatic Matching of ICD-10 codes to Diagnoses in Discharge Letters," *Proceedings of the Workshop on Biomedical Natural Language Processing, Hissar, Bulgaria*, p. 11–18, 2011.

[11] P. Foudeh and N. Salim, "Information Extraction from Handwritten Medical Records and Assigning ICD-10 Codes," in *proceeding of International Workshop of Extraction of Structured Information from Texts in the Biomedical Domain ESIT-BioMed 2010*, Kuching, Malaysia,, July 2010.

[12] «terrier,» 2011. [Çevrimiçi] : http://www.terrier.org/. [1/10/2012 tarihinde erişilmiştir].

Sorumlu Yazarın Adresi

Adil ALPKOÇAK, Dokuz Eylül Üniversitesi, Bilgisayar Mühendisliği Bölümü, Tınaztepe 35160, İzmir