MEDİKAL METİN AYRIŞTIRMA VE SINIFLANDIRMA: SEMPTOMDAN HASTANE BRANŞINA

Ahmet BARDIZ
Galatasaray Üniversitesi – Fen Bilimleri Fakültesi Bilgisayar Mühendisliği Bölümü
34349 Beşiktaş İstanbul, ahmetbardiz@gmail.com

Özetçe

Bu çalışma hastane yönetim sistemlerinde yer alan hekimler tarafından girilmiş hastalara ait eski kayıtları referans alarak hastaların şikayetleriyle ilgilenebilecek hastane branşını makine öğrenmesi yöntemleriyle tespit etmeyi kapsamaktadır. Fakat sınıflandırmadan önce karşımıza çıkan ilk zorluk Türkçe medikal metin verisi üzerinde çalışıyor olmaktır. Türkçe' nin sondan eklemeli bir dil oması ve kendine özgü dil bilgisi kurallarının olması klasik metin ayrıştırma yöntemlerini kullanmaya engel olmaktadır. Bu nedenle sınıflandırma performansını artırmak için Türkçe medikal kelime veya kelime gruplarını ayrıştıran ve semantik olarak anlamladıran bir Türkçe doğal dil işleme servisi geliştirilmiştir. Bu servis üzerinde işlenip anlamlandırılan medikal metin verileri karar destek makinesi ve Bayesian makine öğrenmesi algoritmaları kullanılarak sınıflandırılmış ve yapılan testler sonucu en yüksek doğruluk oranı %97.7 olarak hesaplanmıştır. Ayrıca bu medikal Türkçe doğal dil işleme ve sınıflandırma çalışması paketlenerek herhangi bir sistem veya web ortamına entegre olabilecek şekilde servis olarak yayınlanmıştır. Böyle bir servisin hastane randevu süreçlerinde kullanılmasıyla, hastaların şikayetleriyle örtüşen daha doğru ve özelleşmiş bir hastane branşına yönlendirilmesi hedeflenmiştir.

1.Giriş

Muayene olan hastalar randevularını genel olarak çağrı merkezini arayarak, online randevu web veya mobil randevu kanallarını kullanarak veya doğrundan hastaneye giderek randevu almaktadır. Bir hastanın hangi hastane branşına gidip tedavi olması gerektiğine ise hasta kendi inisiyatifiyle karar verebilirken, çağrı merkezi veya hasta kabul personelleri ya da hekimler tarafından da yönlendirilebilmektedir. Fakat hasta kendi inisiyatifiyle ya da yeterince uzman olmayan personeller tarafından yönlendirildiklerinde yanlış hastane branşından randevu alabilmektedir. Bu durum ya hastanın doğru olmayan branşlarda kalitesiz hizmet almasına sebep olmakta ya da bir daha randevu almak zorunda kalarak hastanın ve hekimin zaman kaybına, kurum ve hastanın da maddi kaybına sebep olabilmektedir.
Türkiye Sağlık Bakanlığı' nın verilerine göre 2016 yılı ilk 6 aylık dönemde Türkiye'deki yaklaşık muayene sayıları şu şekildedir:

- Sağlık Bakanlığı: 170 Milyon
- Özel Hastaneler: 34 Milyon
- Üniversite Hastaneleri: 19 Milyon [1]

Kanallara göre Türkiye randevu oranları ise aşağıdaki şekildedir:

- %62.67 - ALO 182
- %31,10 – MHRS (Merkezi Hastane Randevu Sistemi) Web
- %4.11 - Mobil Uygulama
- %2,12 – Diğer Kanallar [2]

Bu yüksek muayene adetleri ve kanal dağılımları göz önüne alındığında hastalara randevu verilirken çağrı merkezi maliyetlerini düşürmek ve randevu sürecini hızlandırmak için online kanalların teşvik edilmesi gerekmektedir. Fakat online kanallardaki artışla beraber, hastanın randevu alırken şikayetlerini belirtip danışabileceği bir mekanizmanın olmayışı, hastaların yanlış branştan randevu alma oranlarının artmasına sebep olacaktır. Bu nedenle yanlış branştan randevu oranlarını düşürmek için online kanallara ve hastalara branş yönlendirmesi yapan personellerin kullandığı sistemlere hastanın şikayetlerine göre hastane branşı önerisinde bulunan servisler entegre edilebilir.

Hasta veya ilgili personellerden serbest metinle alınacak olan hasta şikayetlerine en olası hastane branşını önerebilecek, entegre edilebilir bir servis geliştirmek üzere bu çalışma yapılmıştır. Bu çalışma kapsamında daha öncesinde hastane yönetimi sistemlerinde hekimlerce kapatılan hasta kayıtları güvenilir bir veri kaynağı olarak kabul edilerek, bu metinsel veri üzerinde Türkçe doğal dil işleme ve sınıflandırma çalışmaları yapılmıştır.

1. Medikal Metin Ayrıştırma

1.1 Türkçe Doğal Dil İşleme

Metin verisi ile sınıflandırılmak istendiğinde doğruluk oranlarını artırmak için metin ön işleme yapmak gerekir. Bu kapsamda anlamlı anahtar kelime veya kelime gruplarını tespit edip, kelimeler gereksiz eklerinden temizledikten sonra bir model oluşturulur. Türkçe'nin sondan eklemeli zengin bir dil olması ve kendine has dilbilgisi kuralları nedeniyle bu aksiyonları almak zorlaşmaktadır [3]. Özellikle İngilizce' deki çok daha basit dilbilgisi kuralları, hazır kütüphane ve veri setleri dolayısıyla metin verisi ayrıştırmak Türkçe'ye kıyasla çok daha kolay ve performanslı yapılabilmektedir [4].

Hastaların şikayetlerini içeren medikal verileri ayrıştırırken kelimeleri gereksiz eklerinden arındırıp anlamsal bütünlük içeren kelime gruplarını tespit etmek gerekmektedir. Fakat kelimeleri eklerinden arındırırken kelimenin karakteristiğini de kaybetmemek gerekir. Örneğin 'kansızlıktan' kelimesini ayrıştırırken kelimenin köküne inersek aslında esas yakalamamız gereken 'kansızlık' semptomunu kaybederiz. Bu nedenle sadece çoğul eki, sahiplik eki, zaman ekleri gibi kelimenin semantik yapısını bozmayacak gereksiz eklerden arındırılmış en kapsamlı halini ayrıştıran bir servise ihtiyaç duyulmaktadır.

1.2 Medikal Metin Ayrıştırma Servisi

Kelimeler gereksiz eklerinden arındırıldıktan sonra hastane branşı atamasına etki edebilecek, şikayet içerisinde geçen değerli kelime veya kelime grupları semantik olarak işaretlenmeli ve sınıflandırılmalıdır. Örneğin 'baş ağrısı' kelime grubu incelendiğinde 'baş' kelimesinin bir ORGAN 'ağrı' kelimesinin ise bir SEMPTOM olduğu etiketlenmelidir. Ayrıca 'baş' ve 'ağrı' kelimeleri bir araya geldiğinde daha özel bir semptomu işaret ettiği bilinirse, hastane branşı önerme aşamasında büyük bir katma değer sağlayacaktır.

Hastanın tedavi olabileceği hastane branşını belirleyen yegane unsur semptomları değildir. Sahip olduğu semptomların periyotları, durumları gibi özellikler de daha kesin öneri yapmayı sağlayabilir. Örneğin 'dokununca göğüs ağrısı' ve 'içten göğüs ağrısı' şeklindeki farklı iki şikayette temelde 'göğüs ağrısı' semptomu bulunmaktadır. Fakat bu semptomun farklı özellikleri farklı hastane branşlarına işaret ediyor olabilir. Bu nedenle semptomları özelleştiren bu tamlayanlar semptomlardan ayrı olarak değerlendirilmemeli ve tümü anlamlı bir kelime grubu olarak işaretlenmelidir.

Şekil 1. Otomatik kelime grubu oluşturma

Türkçe dilbilgisi kuralları kodlanarak genel bir Türkçe doğak dil işleme kütüphanesi oluşturulabilir fakat medikal içeriği anlamak için ayrıştırma servisini medikal sözlüklerle beslemek gerekir. Örneğin literatürde yer alan semptomlar, bu semptomların hem hekim hem de halk arasındaki olası söylemleri, bu semptomların periyotları ve durumları gibi bilgiler sisteme tanıtılmalıdır. Ayrıca yanlış yazımlarını, eş anlamlı kelimeleri veya farklı telaffuzları yakalama yetkinlikleri de ayrıştırıcıya tanımlanmalıdır. Sonrasında ayrıştıcı bu sözlük ve kuralları kullanarak gerekirse yanlış yazımları düzeltip, kelime gruplarını yakalayarak semantik olarak etiketleyebilmelidir.

```
"tokens":
{
"token": "şiddetli kafa ağrsından",
"tokenType": "SPECIFIC_SYMPTOM",
"normalize": "şiddetli baş ağrısı",
"meta": {
        "synonym": true,
        "phonetic": false,
        "fuzzy": true
 }
```

Şekil 2. Türkçe medikal metin ayrıştırma servisi örnek çıktısı

Java yazılım dili tabanlı yazılan Türkçe medikal metin ayrıştırma servisinin yukarıdaki json formatındaki örnek çıktısında servise gelen serbest bir metindeki yanlış yazım düzeltilmiş, bir kelime daha yaygın kullanılan eş anlamlısı ile değiştirilmiş, bu metinde özel bir semptom olduğu tespit edilmiş ve gereksiz eklerinden arındırılarak standartlaştırılmıştır.

1.3 Medikal Metin Sınıflandırma

Medikal metin ayrıştırma servisinden geçen veriler etiketlenip anlamdırıldıktan sonra verilerin sınıflandırılması aşamasına geçilebilir. Hastane bilgi yönetimi sisteminde (HBYS) uzman hekimlerce kapatılmış hasta kayıtları referans alınarak yeni gelen bir hastanın şikayetlerine en uygun hastane branşını önerecek eğitilmiş bir model geliştirilmek istenmektedir. Bu bağlamda metin sınıflandırmada öne çıkan iki gözetimli öğrenme yöntemi olan Destek Vektör Makineleri (DVM) ve Çok Terimli Naive Bayes (NB) algoritmaları kıyaslanmıştır [5].

1.4 Medikal Örnek Veri Kümesi

Bu çalışmada kullanılan örnek veri kümesi anonimleştirilerek Türkiye'deki bir üniversite hastanesinin HBYS sisteminden çekilmiştir. Her bir kayıt branşında uzman hekimlerce hastanın şikayet ve hikayesi HBYS sistemine kayıt edilerek oluşturulmuştur. Bu örnek veri kümesindeki kayıtlar aşağıdaki özelliklere sahiptir:

- 10 Farklı Hastane Branşı
  - Fizik Tedavi ve Rehabilitasyon
  - Dermatoloji
  - Gastroenteroloji
  - Kadın Hastalıkları
  - Kardiyoloji
  - Kulak Burun Boğaz
  - Ortopedi ve Travmatoloji
  - Plastik Cerrahi
  - Psikiyatri

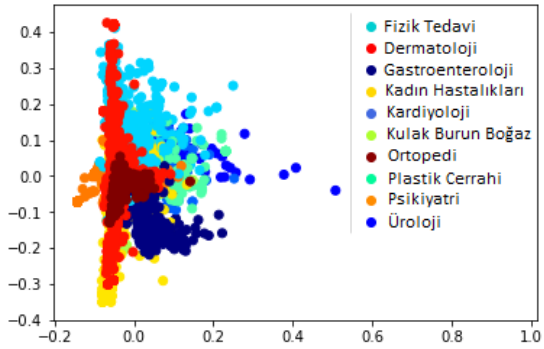
- Üroloji

- Her bir hastane branşı için yaklaşık 10.000 doküman (hasta kaydı)
- Her bir doküman iki özelliğe sahiptir. Biri hastanın şikayet ve hikayesini içeriyorken diğeri bu bilgiler ışığında hekim tarafından atanan hastane branşıdır.

Bu çalışma kapsamında toplamda yaklaşık 100.000 kayıt içeren bu veri kümesi öğretme ve test aşamalarında kullanılmak üzere iki parçaya bölünmüştür. Kayıtların %90'ı öğretmek için, geri kalan %10'u da test aşamasında kullanılmak üzere ayrılmıştır. Test için kullanılacak veri kümesi hiçbir şekilde öğretme aşamasında kullanılmamıştır.

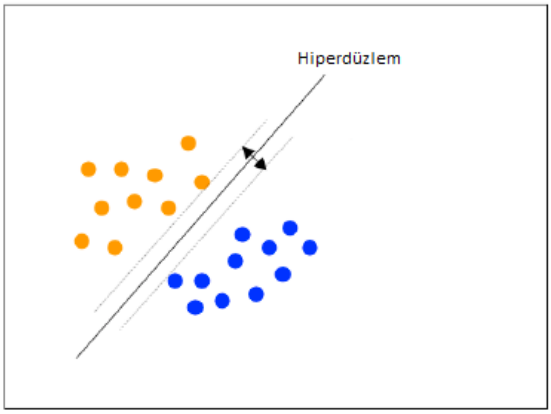
2. DVM ile Metin Sınıflandırma

DVM metin sınıflandırmada yaygın olarak kullanılan en kuvvetli yöntemlerden biridir. Metinsel verimizi vektörel ifadelere dönüştürdükten sonra bir düzleme yerleştirilerek hangi sınıfa ait oldukları işaretlenir.



Şekil 3. İki boyutlu düzlemde vektörel gösterim

Sonrasında birbirinden en uzak olacak şekilde hiperdüzlemler kullanılarak vektör kümeleri birbirinden sınırlandırılır. Böylece her bir sınıfın düzlem üzerinde kendine ait bir bölgesi oluşur. Yeni gelen bir doküman da bu düzlemdeki yerine göre, denk geldiği bölgenin sınırlarına sahip sınıfa ait olur.



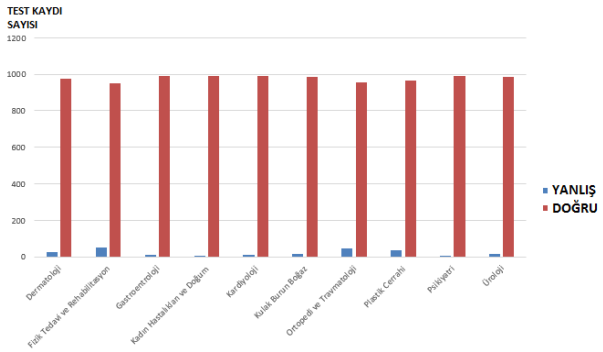
Şekil 4. DVM hiperdüzlemi [6]

İlk yaklaşım olarak eğitim için ayırdığımız veri kümesini kullanarak her bir dokümanı Türkçe medikal metin ayrıştıma servisimizden geçirerek anlamlı ve değerli kelime veya kelime gruplarını belirleyerek servis çıktısını ile eğitilmiş bir model oluştururuz. Daha sonrasında test için ayırdığımız veri setini kullanarak eğittiğimiz modelin tahmin ettiği branşlarla esas veri setinde mevcut olan hastane branşları karşılaştırılarak oluşturulan modelin performansı ölçülür. Bu çalışmada python yazılım dilinde temelde scikit-learn kütüphanesinde yer alan doğrusal DVM çekirdeği kullanılmıştır [7]. Sigmoid, polinomial, gaussian gibi farklı çekirdekler kullanılarak performansları değerlendirilebilir.

2.1 DVM Test Sonuçları

Eğittiğimiz DVM modelinin performansını ölçebilmemiz için daha öncesinde ayırdığımız test verisini kullanırız. Bu test verisinde hekim tarafından atanmış hastane branşları doğru kabul edilir ve modelin tahmin ettiği çıktı ile karşılaştırılır. Daha yüksek oranda bir eşleşme sağlamak için eriler eğitilirken yapıldığı gibi test verileri de Türkçe medikal metin ayrıştırma servisinden geçirilir.

Bu çalışmada kullandığımız örnek veri kümesi için başarı oranı %97.7 çıkmıştır. Başarı oranının bu kadar yüksek çıkmasına, eğitim ve test verilerinin sadece hekimler tarafından sisteme girilmiş olması neden olmuş olabilir. Çünkü uzman hekimlerin kendi literatürlerinde kulllandıkları ortak terim, kısaltma veya pratik bilgiler yer alabilir. Hastalar şikayetlerini ve hikayelerini kendi dillerinden yazmaya başladıkları zaman başarı oranı muhtemelen düşecektir. Ayrıca örnek veride yer alan 10 adet klinik için test yapılmıştır. Tüm klinikleri kapsayacak bir çalışma yapıldığında başarı oranının düşmesi muhtemeldir.



Şekil 5. Kliniklere göre DVM başarı oranları

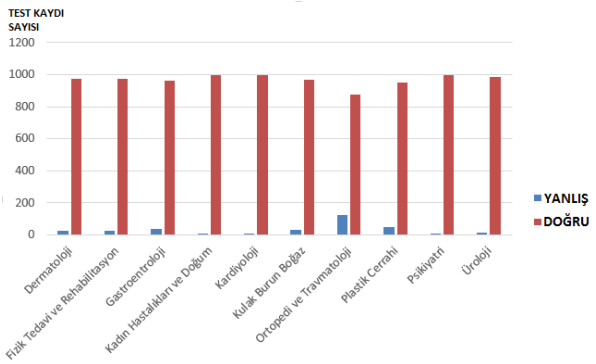
Eğitilen DVM sınıflandırıcının kliniklere göre doğru ve yanlış tahminlerinin dağılımı yukarıdaki grafikte görülebilmektedir. Grafik dikkatli incelendiğinde en fazla hata payı 'Fizik Tedavi ve Rehabilitasyon' ve 'Ortopedi ve Travmatoloji' branşlarında yer almaktadır. Bu iki branşın uzmanlık konuları birbirine yakın olduğu için DVM tahminleri birbirine geçebilmektedir. Bu da bu iki poliklinik için başarı oranı düşürmektedir.

2.1.1 Multinomial Naïve Bayes Sınıflandırma

Naive Bayes sınıflandırma kolay uyarlanması , hızlı ve tutarlı olması sebebiyle bir çok alandan ilk tercih edilen yöntemlerden biridir. Semptom teşhisinde ve tedavi süreçlerinde karar mekanizması olarak kullanıldığı örnekler bulunmaktadır [8]. Biz de bu çalışma da DVM ile kıyaslamak üzere çok terimli NB tercih edeceğiz. Çünkü çok terimli NB özellikle metin sınıflandırmada, kelimelerin çoklu tekrarları önem arz ettiğinde kullanılır [9].

2.1.2 NB test sonuçları

DVM model oluşturulurken kullanılan eğitim ve test verisinin aynısı NB ile modellenip test edildiği zaman başarı oranı %96.7 çıkmaktadır. DVM ile karşılaştırıldığında sadece %1' lik bir düşüş olmaktadır.



Şekil 6. Kliniklere göre NB başarı oranları

Test sonuçlarında DVM çok az da olsa daha başarılı gözükmektedir. Test sonuçlarına paralel olarak metin sınıflandırmada genel olarak DVM, NB' ye oranla daha yüksek performans gösterdiği kabul edilmektedir. NB veri hakkında herhangi bir ön bilgiye sahip olmadan basit ve tatmin edici bir sınıflandırma sağlarken, DVM daha özel ve ayrıştırıcı bir süreç uygulayarak metin sınıflandırmada çok daha iyi bir performans sergiler [10].

2.1.3 Kesinlik, Hassasiyet ve F-Skorları

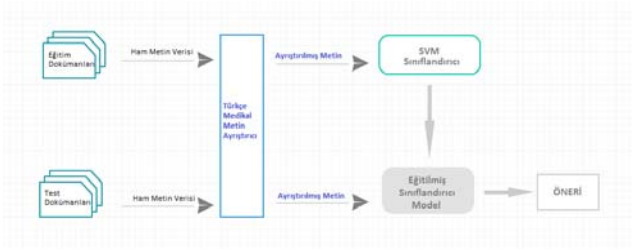
Yaptığımız çalışma sonucunda DVM ve NB için çıkan ortalama kesinlik, hassasiyet ve f-skor oranları aşağıda listelenmiştir.

- **DVM**
  - Kesinlik: 0.979
  - Hassasiyet: 0.979
  - F-skoru: 0.979
- **NB**
  - Kesinlik: 0.968
  - Hassasiyet: 0.967
  - F-skoru: 0.967

Kesinlik, hassasiyet ve f-skor oranları test sonuçlarında çıkan doğruluk oranlarına paralel olarak DVM de küçük bir farkla daha yüksek çıkmıştır. Training ve test verileri homojen olarak dağıtıldığı için bu oranlarda sürpriz bir fark oluşmamıştır.

2.2 Rest-API Servis

Bu yapılan çalışmada kullanılan Türkçe medikal metin ayrıştırma servisi ve performansı daha iyi olduğu için DVM sınıflandırma servisi paketlenerek bir çözüm olarak yayınlanmıştır. Bu mimari tüm klinikleri içeren daha kapsamlı bir veri kümesi ile eğitilerek ve Türkçe medikal metin ayrıştırıcı zenginleştirilerek gerçek ortam uygulamalarında hizmete açılabilir. Hastanelerin online randevu sistemlerine entegre olarak şikayetlerini girecek olan hastayı doğru branşa yönlendirebileceği gibi çağrı merkezi ve hasta kabul personellerinin kullandığı sistemlere de entegre olarak personellerin hastanın şikayetlerini alarak hastayı daha doğru bir branşa yönlendirmeleri sağlanabilir.



Şekil 7. API servis mimarisi

3. Sonuç

Bu çalışmada geliştirilen ve %97.7 başarı oranı elde edilen Türkçe ve DVM tabanlı hastane branşı öneri servisi, hastaların şikayetleriyle örtüşen daha doğru branşlardan hizmet almasını sağlayacaktır. Böylelikle hastalar tarafından online kanallardan alınan veya çağrı merkezi personelleri tarafından verilen yanlış randevular nedeniyle oluşan mağduriyet, hastanın ve hekimin zaman kaybı ve hastanelerin maddi kaybı azalacaktır.

Türkçe medikal sözlük kapsamı genişletildiğinde ve detaylandırıldığında branş öneri sisteminin performansı daha da artacaktır. Özellikle hekimleri ve hastaları ortak bir noktada birleştirmek için semptomların halk dilindeki söylemleri medikal sözlüğe tanımlanabilir. Aynı zamanda bu sözlüğün internetten veya HBYS verilerinden otomatik olarak genişletilmesi sağlanabilir. Medikal sözlükteki kelime veya kelime gruplarının yakınlıklarını derecelendirecek bir medikal semantik sözlük oluşturarak öneri sistemine dolaylı olarak etki etmesi sağlanabilir. Ayrıca halihazırda kullandığımız 'bag of words' yaklaşımına ek olarak cümlelerin de öğelerine ayrılıp öğelerin birbirleriyle olan ilişkileri de arama sonuçlarına ve modellemeye yansıtılabilir [11].

KAYNAKLAR

[1] Türkiye Sağlık Bakanlığı 2016 6 Aylık Genel Hizmet Bilgileri Tablosu, http://rapor.saglik.gov.tr/istatistik/rapor/

[2] H. Ömer TONTUŞ: Türkiye'de Sağlık Hizmetine Kolay Erişim Merkezi Hekim Randevu Sistemi (MHRS), pp. 2-6 2015.

[3] Dilek Z. Hakkani-Tür, Kemal Oflazer, Gökhan Tür: Statistical Morphological Disambiguation for Agglutinative Languages, pp. 1-3, 2000.
[4] Zeynep Boynukalın: Emotion Analysis of Turkish Texts by Using Machine Learning Methods, pp. 3, 2012.
[5] Jason D. M. Rennie, Ryan Rifkin: Improving Multiclass Text Classification with the Support Vector Machine, pp. 6-8, 2002.
[6] Doina Caragea, Dianne Cook, Vasant Honavar Visual Methods for Examining Support Vector Machine Results, with Applications to Gene Expression Data Analysis, pp. 4, 2005.
[7] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion: Scikit-learn: Machine Learning in Python, 2011.
[8] Ashraf M. Kibriya, Eibe Frank, Bernhard Pfahringer, Geoffrey Holmes: Multinomial Naïve Bayes for Text Categorization Revisited, pp.1, 2005.
[9] Joanna Kazmierska, Julian Malicki: Application of the naive bayesian classifier to optimize treatment decisions. Radiotherapy and Oncology, pp. 211–216, 2008.
[10] Thorsten Joachims: Text categorization with support vector machines: learning with many relevant features, pp. 137-142, 1998.
[11] Simon Tong, Daphne Koller: Support Vector Machine Active Learning with Applications to Text Classification, pp. 62, 2001.