## ÖZET

Bu çalışmada, Türkçe haber metinleri üzerinden sınıflandırma yöntemleri kullanarak özetleme yapan bir metin özetleme sistemi sunmaktayız. Sistemimizde, haber metnini en iyi temsil edecek şekilde cümleler seçilir. Sistemde kullanılan sınıflandırmalar için haber metnini temsil etmeye en uygun cümlelerin özellikleri olarak cümle konumu, anlatım ifadeleri, varlık isimleri, kelime sıklığı ve başlık benzerliği kullanıldı. Sınıflandırmalarda ve değerlendirmelerde kullanılan etiketler, referans özet olarak düşünülen haber açıklamalarından edinildi. Türkçe metin özetleme üzerine yapılan sayılı çalışmalardan biri olarak, başarılı sonuçlar elde edilmiştir.


## GİRİŞ

İnternetin yaygınlaşmasıyla beraber haber kaynaklarına ulaşmak da inanılmaz derecede kolaylaştı. Dijital ortamın hızlı ve masrafsız imkanlar sağlamasından dolayı, haber kaynakları istedikleri sayıda ve uzunlukta haberler yayınlayabilmektedir. Özelikle sosyal medyanın da hayatımızın bir parçasıyla beraber, haber kaynakları bu kanal üzerinden de bizlere sürekli olarak haber sunmaktadır. Dolayısıyla dijital ortam kullanıcıları olarak, onlarca haber kaynağından yüzlerce ve sayfalarca haberlere anında erişebilmekteyiz. Gündemi takip edebilmek için her gün onlarca haber okuyoruz, ama çok zaman harcamamıza rağmen birçok haberi gözden kaçırdığımız oluyor. Aslında kısa ve net ifadeler okurlar için yeterli olacakken, uzun detaylar, gereksiz bilgiler, alakasız haberler ile vaktimizi harcamaktayız. Kimi zaman aynı bilgiyi tekrarlayan, kimi zaman ticari kaygılarla uzun haberler sunan kaynaklar ise okurlar açısından gerçekten zaman kaybı olabilmektedir. Yani, günümüzde her ne kadar rahatça haber kaynaklarına ulaşsak da yine rahatlıkla zamanımızdan olabiliyoruz. Bu durumda, biz de projemizde bir Türkçe haber özetleme aracı geliştirip dijital haber okurlarına haber özetlerini verimli bir şekilde sunmayı hedeflemekteyiz.

Toplumumuzun internet kullanımı üzerine yapılan araştırmayı göz önüne aldığımızda [1], internet ve sosyal medya kullanımının hızla arttığını ve geleneksel medya takibinin azaldığı gözlemlenmektedir. Bu yüzden böyle bir geliştirmeye ihtiyaç da artmaktadır. Her ne kadar sosyal medyada geçen zamana kıyasla daha az olsa da yine de haber sitelerinde geçen zaman hatırı sayılır derecelerdedir. Bunun gibi sebepler çalışmakta olduğumuz projenin motivasyonunu artırmaktadır.

Bu çalışmada, "Verilen bir haber metnini en iyi özetleyen cümleleri nasıl seçebiliriz" araştırma sorusuna eğildik. Haber özetlemesi de genel olarak metin özeti kapsamına girmektedir ve bilgi erişiminin kolaylaşıp kaynak metinlerin hızla çoğalmasıyla özetleme sistemleri üzerine birçok çalışma yapılmıştır [3,4,5,6]. Yapılan çalışmalar göz önüne alındığında temel olarak iki çeşit metin özetleme yaklaşımı vardır: cümle seçerek özetleme ve yorumlayarak özetleme. Yorumlayarak özetlemede, kaynak metindeki ifadeler mantıklı bir şekilde kısaltılarak tekrar yazılmaya çalışılır. Cümle seçerek özetlemede ise, kaynak metindeki önemli cümleleri, istatistiksel metotlar, sezgisel çıkarımlar veya bunların ikisinin birleşimiyle seçerek özetleme yapılır. Özeti oluşturan cümleler, yorumlamada yeniden kurulmuşken, cümle seçmede kaynak metinden seçilmektedir.

Bizim projemizde cümle seçerek özetleme yöntemini kullanılır. Çeşitli kaynaklardan seçtiğimiz haber metinleri ve açıklamaları edinilir. Öncelikle haber metinlerimiz çeşitli ön işlemlerden geçirilir: cümlelere ayırma, normalize etme, kelimeleri eklerinden ayırma, gereksiz kelimelerden arındırma ve büyük-küçük harf ayarlama. Referans özetlerimiz kullanılarak haber metinlerindeki cümleler etiketlenir. Burada kosinüs benzerliği ile en referansa en benzeyen cümleler işaretlenir. Bir haber metinden özet çıkarmak için en uygun cümleleri seçerken sınıflandırma yani güdümlü öğrenme yöntemi kullanılır. Sınıflandırma işlemini yaparken, terim sıklıklarına çeşitli öznitelik seçimleri yapılır. Sınıflandırıcılar olarak farklı parametrelerle SVM ve Rassal Ormanlar kullanılır. Sistemimizin edindiği sonuçlar yani özetler ise ROUGE [2] kullanılarak olması gereken özet ile kıyaslanır. Elde edeceğimiz sonuçlara göre farklı parametreler veya araçlar kullanarak geliştirme yapabiliriz.

## İLGİLİ ÇALIŞMALAR

Uzundere ve ekibinin [3] çalışması, eski olmasına rağmen, bizim projemize en benzer olan projedir. Cümle seçerek özetleme yöntemi kullanılmıştır. Haber metni ön işlemlerden geçtikten sonra; başlık, özel isim, yüksek frekans gibi özelliklere göre cümlelere puan verilir. Sezgisel olarak her özelliğin puan

katsayısı verilmiştir. En yüksek puanlı belirli sayıda cümleler özet için seçilir. Referans olarak insanların cümle seçerek oluşturduğu özetler kullanılmıştır. Sonuç olarak %55 e yakın oranlarda benzer özetler elde edilmiştir. Kutlu ve ekibinin [4] çalışmasında ise öncekinden daha güncel yöntemler çalışılmış olup yine cümle seçerek Türkçe genel özetleme yapılmıştır. Farklı özelliklere; cümle konumu, başlık benzerliği vb., göre puanlama fonksiyonları kullanılmıştır. Bunlardan anahtar cümle (key phrase) özelliği bu projeye özel kullanılmıştır. Makine öğrenmesi yöntemleriyle kullanılan özelliklere ağırlıklar sağlanmaktadır. Ağırlıkları eğitim veri kümesinden en yüksek puanlamayı yapacak şekilde öğrenmektedir. Elde edilen özetler insan yapımı özetler ile ROUGE kullanılarak karşılaştırılmıştır. Bizde ise haber açıklamaları referans özetler olarak kullanılır.

Keneshloo ve ekibinin [5] çalışmasında, cümle seçerek özetleme için kosinüs benzerliğini kullanmıştır. Metindeki her cümlenin birbiriyle arasındaki benzerlikler hesaplanmış olup daha sonra her cümle için ortalama alınmıştır. Benzerlik ortalaması en yüksek olan cümleler özet için seçilmiştir. Biz etiketlemede benzer yöntem uygularız. Yine benzer yöntemle Dutta ve ekibinin [6] çalışmasında, ön işleme uğrayan metin cümlelerinin, birbiri arasındaki kosinüs benzerlikleri elde edilir. Her cümle çizginin düğümleri olmak üzere, bir benzerlik çizgisi ve benzerlik matriksi elde edilir. Çizginin kenarları ise cümleler arasındaki benzerliğine göre ağırlık kazanır. Her düğüm için kümeleme katsayısı çıkarılır. Çizgi ve matriks kullanılarak Infomap kümelemesi yapılır. Her kümeden, kümeleme katsayısı, kümelerin maksimum ortalama kümeleme katsayısından fazla olan cümleler seçilir. İnsanlar tarafından yapılan özetler referans olarak kullanılarak, farklı ROUGE yöntemleri ile sonuçların değerlenmesi yapılmıştır. ROUGE-1 ile 0.5'den fazla duyarlılık, kesinlik ve f1 değerleri elde edilmiştir.

Khan ve ekibinin [7] çalışmasında, k ortalamalar ve TF-IDF yöntemi kullanılarak cümle seçerek özetleme çalışması yapılmıştır. Silhouette ve elbow yöntemleriyle K değeri belirlenmiştir. Özet sonuçlarını karşılaştırmak için çevrimiçi bir metin özetleme sitesi kullanılmış ve sonrasında BLEU adı verilen bir yöntemle sonuçların değerlendirmesi yapılmıştır. Belirtildiğine göre ilerleyen çalışmalarda daha detaylı ön işlemler, gereksiz cümlelerin atılması gibi, yapacaklar. Verimli sonuçlar alınmıştır. Bir diğer kümeleme çalışması olarak Alguliyev ve ekibinin [8] çalışmasında, iki adımda kümeleme ve optimizasyon yöntemlerini birlikte kullanarak cümle seçerek özetleme üzerine çalışma yaptılar. Optimize ederken; kapsama, özet kaynak metnin ana konusunu farklı açılardan kapsamalı; çeşitlilik, özette aynı bilgiyi tekrar eden cümleler olmamalı; uzunluk, özet ve seçilen her cümle belirli uzunlukta olmalı. Bizim çözümümüz de benzer şekilde K ortalamalar kümeleme kullanarak benzer cümleleri kümeleyecek ve en ilgili kümelerden birer cümleler seçilecek. Elde edilen sonuçlar ROUGE ile değerlendirilmiştir. İlgili çalışmalarla, 10 farklı, kıyaslandığında gerçekten başarılı sonuçlar ortaya çıkardığı gözlemlenmiştir. Başka bir kümeleme kullanan Mandal ve ekibinin [9] çalışmasında, parçalı küme zekası (Particle Swarm Optimization (PSO)), metin özetleme için kullanılmıştır. Açık kaynak kodlu araç olarak SentiWordNet, kullanılarak duygusal analiz (SA) yapılmıştır. PSO'da kullanmak için, duygusallık puanı ile uygunluk değeri hesaplanmıştır. PSO ile kümeleme yapılmıştır. Önce küme sayısı belirlenip, daha sonra cümle kümelemesi için Otomatik Kitle Bölümü (APP) uygulanmıştır. ROUGE ile sonuçlar değerlendirilmiştir. Sonuç olarak PSO ve SA ile özetleme yapılmıştır. Çalışmalarda güdümsüz öğrenme uygulanmıştır, bizde ise güdümlü uygulanır.

Yadav ve Shah[10] çalışmalarında, diğer çalışmalara benzer şekilde metinler ön işlemden geçirilir. Cümle seçme yöntemiyle metin özetlemesi için Fuzzy mantığı kullanılmıştır. Başlık benzerliği, uzunluk vs. kullanılarak özellikler matriksi oluşturulmuş ve Fuzzyde kullanılmıştır. Fuzzy ile en uygun cümleler seçilerek özet elde edilmiştir. Sonuç olarak başarılı olarak değerlendilen bir çalışmadır.

Sadece ingilizceden farklı olarak Abujar ve ekibi [11]nin çalışmasında, İngilizce ve Bengal Dili üzerine blog, haber vb. türlerde çalışma yapılmıştır. Sözcüksel, anlamsal ve sözdizimsel analizler yapıldığı belirtilmekte. Kelime, cümle ve sıralama benzerlikleri ölçülmektedir. Benzerlik hesaplamaları için, mesafe ölçümü, Levenshtein uzaklığı, Wu ve Palmer(WO) ölçümü ve Lin ölçümü gibi yöntemler kullanıldı. Kullanılan yöntemlerin aynı örnek üzerinde performansı gösterilmiştir. Daha iyi sonuçlar için geri-izleme gibi yöntemlerin daha iyi olabileceğini belirtilmektedir. Benzerlikleri bulmak konusunda farklı yöntemlerin kullanılması üzerine başarılı bir çalışmadır.

## ÖNERİLEN YÖNTEM

Özetleme sistemlerinde genel sorunlardan bir tanesi, özellikle sınıflandırma yapılacaksa, veri kümesi oluşturulmasıdır. Özetleme sistemlerinin temel sorunu referans özetlerdir. Özetlenecek metinlerin edinilmesinden ziyade özetlerinin edinilmesi ise daha zorlu bir süreçtir. Bu bölümde, haber metninden cümle seçerek özetleme yapan sistemimizi belirtilen önemli sorunlara da çözümler oluşturarak açıklayacağız.

Sistemimizdeki amacımız, sınıflandırma yöntemlerini kullanarak haber metninden en uygun cümleleri seçerek özet oluşturmaktır. Sistemimiz için girdimiz Türkçe haber metni olup, çıktımız ise girdi metninden seçilmiş özet için uygun olan cümleler olacaktır. Örneğin, Figure 1'de haber metninden özet çıkarılacaktır. İlgili haber açıklaması da Figure 2'dedir, referans özet olarak tanımlanmakta, yani açıklama haber metninin en ideal özetlenmiş hali olarak varsayılır. Her ikisi de aynı haber sayfasından edinilmiştir [12]. Bu örnek metinden özet çıkarmak istendiğinde seçilecek cümlelerin Figure 3'deki gibi olması muhtemeldir. Geliştirilecek olan sistemimizin bu şekilde özet üretmesi beklenmektedir.

İzmir'in Tire ilçesine bağlı Işıklı Mahallesi Karatepe mevkiinde çıktı.
Öğleden sonra makilik alanda başlayan yangın, rüzgarın da etkisiyle büyüdü.
Kısa sürede olay yerine Tire Orman İşletme Şefliği, Bayındır Orman İşletme Şefliğinden ekipler sevk edildi. Bu arada yangına bir helikopterde havadan müdahale etti.
Öte yandan, yangının büyümemesi için vatandaşlarda canla ve başla mücadele verdi.
Vatandaşlar, kendi su tankerleriyle yangına müdahale ederek söndürmeye çalıştı.
Yangının yerleşim yerlerine yakın olması nedeniyle ekipler çalışmalarını yoğun olarak sürdürüyor.

*Figure 1*

**İzmir'in Tire ilçesinde makilik alanda yangın çıktı. Yerleşim yerlerine yakın olan yangına karadan ve havadan müdahale ediliyor.**

*Figure 2*

İzmir'in Tire ilçesine bağlı Işıklı Mahallesi Karatepe mevkiinde çıktı. Yangının yerleşim yerlerine yakın olması nedeniyle ekipler çalışmalarını yoğun olarak sürdürüyor.

*Figure 3*

Sistemimiz ise cümle seçerek özetleme sürecini sınıflandırma ile yapacağı gereken cümleleri etiketleme işlemi önemli bir yere sahiptir. Etiketleme bölümünde bu konu detaylıca incelenmektedir.

Öncelikler veri kümemiz haber metni ve açıklamalarıyla oluşturulur. Haber metinleri ve açıklamaları, cümlelere ayrılıp ön işlemlerden geçirilir. Açıklamalar, metin cümlelerini etiketlemek için kullanılır. Etiketlenen cümleler, SVM gibi metin sınıflandırmada iyi olan güdümlü sınıflandırma yöntemleri kullanılarak sınıflandırıcılar oluşturularak özetleme sistemi geliştirilmiş. Veri kümesi, ön işlem, etiketleme ve özetleme alt başlıklarda açıklanmaktadır.

1. Veri kümesinin haber metinleri ve açıklamaları edinerek oluşturulması
2. Haber metinlerinin ve açıklamalarının ön işlemden geçirilerek hazırlanması
3. Haber metin cümlelerinin etiketlenmesi, özet için uygun cümleler
4. Sınıflandırma için öz niteliklerin çıkarılması
5. Veri kümesi üzerinde SVM ve rassal ormanlar sınıflandırıcılarının çalıştırılması

## 1. Önişlem

Sınıflandırma işlemlerine geçmeden önce hem haber metinlerinin hem de haber açıklamalarının belirli metin ön işlemlerinden geçmesi gerekmektedir. Haber metinleri ve açıklamalar Türkçe olduğu için, Türkçe için en uygun NLP araçları seçilmeye çalışıldı. Sınıflandırma cümleler üzerinden yapılacağı için, öncelikle metinler ve açıklamalar cümlelere ayrılır. Haber metinleri düzgün metinler oldukları için nokta(.) ile cümlelere ayrılır. Düşük ihtimal de olsa metindeki tamamen aynı cümleler atılır. Daha sonra, Türkçe için kullanımı kolay bir doğal dil işleme kütüphanesi olan Zemberek [13] kullanılarak cümleler kelimelere ayrılır(tokenization). Normalde metin ön işlemede, metindeki kelimelerin normalizasyonu için yazım hatalarına bakılır. Ancak, haber metinleri ve açıklamaları ilgili haber siteleri editörleri tarafından kontrolden geçtiği için yazım hatalarının pek mümkün olmayacaktır. Cümlelerde özne dışında büyük harfle başlayan kelimeler yine Zemberek kullanılarak varlık ismi(named entity) olanlar seçilir, bu isimler öznitelikler için gereklidir. Daha sonra büyük harfler küçük harfe dönüştürülür, varlık isimlerini önceden aldığımız için farklı anlamalara sahip kelime karışıklığı engellenmiş olur. Sonrasında, Türkçe etkisiz kelimeler (stop words) cümlelerden çıkartılır, çeşitli anonim kaynaklarda mevcut olan kelime listeler kullanılır [14]. Etkisiz kelimeleri çıkartılan cümlelerdeki kelimeler, Türkçe dahil birçok dil için köke indirme sağlayan Snowball API[15] kullanılarak köklerine indirgenir.

Bu işlemlerden sonra haber metni ve açıklamaları cümle listelerine, cümleler ise kelime listeleri halinde hazırlanarak sonraki süreçlere hazır hale gelmiş olur.

## 2. Etiketleme

Özetleme sistemimizde amacımız, haber metninden cümle seçerek özet oluşturmaktır. Özet için uygun olan cümleleri bulmak için sınıflandırma kullanılır. Haber metnindeki cümleler etiketlendirilerek sınıflandırma yapılır. Etiketleme için elimizde bulunan haber açıklaması kullanılır. Bu durumda ikili(binary) bir sınıflandırma söz konusu olmaktadır, yani haber metnindeki cümleler, özet için uygundur veya değildir.

Haber açıklamasında bulunan cümleler doğrudan metin içerisinden seçilmediği için, metin cümlelerinin aynısını açıklamada bulup etiketlemek mümkün değil. Bu yüzden metin cümlelerinin, açıklamadaki cümleler ile olan benzerliği göz önünde bulundurularak etiketleme yapılabilir. Haber metnindeki her cümlenin haber açıklamasında bulunan cümleler ile olan yakınlığı kosinüs benzerliği kullanılarak bulunur. Figure 4[16]'de ise formülü yer almaktadır. Bu şekilde tüm cümlelerin açıklama ile olan benzerlikleri elde edilir. Elde etmek istediğimiz özette asıl metnin uzunluğuna göre belirli oranda cümle sayısı olacağını varsayarak, örneğin metinde toplam 10 cümle varsa 3 cümle olacak şekilde, elde ettiğimiz benzerliklerden gereken sayıda metin cümlesi ilgili olarak işaretlenebilir. Bir cümle birden fazla cümle ile en benzerse cümle sırasına göre seçilir. Mesela, Açıklamada iki cümle ve metinde bu ikiye en yakın cümlelerin aynı olduğu varsayarsak; metindeki cümle, açıklamadaki ilk cümleye yakın olarak

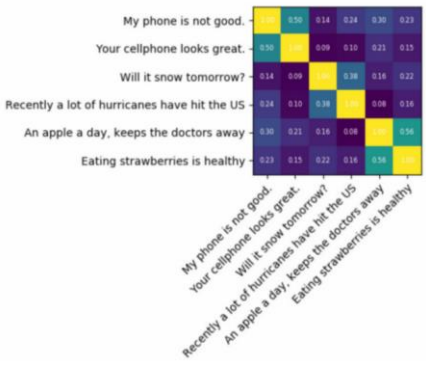


Figure 5

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

Figure 4

işaretlenir ve açıklamadaki diğer cümle için ikinci yakın cümle işaretlenir. Bu durumda belirli bir eşik puanı elde edilmiş olur. Yani metindeki cümleler, eğer açıklama ile benzerliğine göre uygundur (1) veya değildir (0) olarak etiketlenir.

**Açıklamalı [MK1]:** En gelişmiş demek diğer kutuphaneleri kptulemektir ve de bunun en azindan deneysel ispati gerekir.

**Açıklamalı [MK2]:** Cümlelere nasil ayirdiniz?

**Açıklamalı [MK3]:** Ayni cümle birden cok geçiyorsa o zaman sorun olacaktır. Soyle ki, aciklamada ABC cümleleri olsun. Haber metninde de ABCDAEFA olsun. Eger A cümlesi ABC'ye en benzer cümle ise, etiketlerinizde sadece A cümleleri 1 olacak. O yüzden etiketlerde asiri benzer cümleleri 1 yapabilir ama aciklamada 3 cumle varsa 3 tane ayri/farkli cümleyi bulmaniz gerekiyor.

## 3. Öznitelikler Seçimi

Sınıflandırma için gereken bir diğer önemli işlem ise özniteliklerin seçilmesidir. Bizim sistemimizde cümleler sınıflandırılacak olup bir cümlenin hangi özellikler özete seçilebilmek için önemlidir sorusu üzerinden uygun olan özniteliklere karar verilir. Bizim bu konuda seçtiğimiz öznitelikler şu şekildedir:

- Kelime sıklığı(tf): Metni içerisinde yer alan her kelimenin ne kadar sıklıkta geçtiğidir. Üretilecek olan özet, haber metninden seçilme cümleler içereceği için metinde sıklıkla geçen kelimelerin özette bulunması daha muhtemeldir.
- Varlık İsimleri (Named Entity): Varlık isimleri, kişi, kurum, yer vb. varlıklardır. Haber açıklamalarında bu varlıktaki kelimeler sıklıkla geçmektedir
- Alıntı: Bir varlığın doğrudan veya dolaylı söylediği ifadelerdir. Haber metin özetlerine baktığımızda, genelde metin konu olan bir varlığın söylediği sözler önemli yer tutmaktadır. Haber metinlerinde ve açıklamalarında, özellikle gündemde olanlar, genellikle doğrudan veya dolaylı olarak kişilerin ifadelerine sıkça yer verirler. Bu yüzden, bu ifadelere yer verilen cümlelerin özetleme için seçilmesi doğru tercih olacaktır. Doğrudan alıntılar için çift tırnak işareti içeren cümleler kullanılabilir. Dolaylı ifadeler için ise "dedi, söyledi, belirtti, ifade etti vb." gibi kelimeler ve kelime grupları incelenebilir. Bu özelliği varlık isimlerinden ayrı almamızın sebebi, alıntılara doğrudan varlık isimleri olmadan, mesela zamirlerle, de yer verilebilmesidir.
- Konum: Haber metinlerinde özellikle metin başında ve sonunda kullanılan cümleler özet için önemli ifadeler içermektedir. Metin konusunu daha iyi vurgulamak için kullanılan bu cümleler iyi bir tercih olabilir.
- Başlık Benzerliği: Haber başlığı genellikle haber konusunu net ifade eden kelimeler içerir. Haber başlığına yakın olan cümleler metin konusunu daha kapsayan ve özet için uygun olan cümlelerdir.

Yeteri kadar öznitelik çıkartıldıktan sonra, cümleler özniteliklerine göre vektörel olarak ifade edilebilir hale gelmiş olur. Daha önceden yapılan etiketleme ile birlikte artık, sınıflandırma için tüm hazırlıklar yapılmış haldedir. Referans özetlerin yani açıklamaların burada herhangi bir etkisi yoktur, onlar sadece etiketleme için kullanılır.

## 4. Sınıflandırma

Etiketleme işleminin yapılmasıyla birlikte, özet için uygun olan cümleler işaretlenmiş durumdadır. Sınıflandırma kullanarak ise sistemimizin özet için uygun olan cümlelerin hangileri olduğunu seçebilmeyi öğrenmesi beklenmektedir. Yani sınıflandırmanın amacı sistemimize metin verildiğinde içindeki cümlelerden uygun olanlar işaretleyip özet olarak sunabilmesidir.

Sınıflandırma için hazır olan tüm metinler artık birlikte ele alınabilir. Çünkü her cümle vektörel ve etiketlendirilmiş hale geldi. Bu sayede, aynı sınıflandırıcı eğitim kümesi üzerinde tüm haber metinlerinin tüm cümleleri üzerinde çalıştırılabilir.

Cümleler üzerinde ikili sınıflandırma, özet için uygun veya değil, yapacağımız için metin sınıflandırmada sıkça kullanılan SVM ve rassal orman sınıflandırma yöntemlerine başvurulur. Sınıflandırma algoritmaları konusunda hazır geliştirmelere sahip ve oldukça kolay kullanımı olduğundan dolayı Scikit-learn[17] kütüphanesi kullanılır. Belirtilen sınıflandırma yöntemleri farklı hiper parametrelerle, mesela SVM için kernel fonksiyonu, en başarılı sonuçları elde edecek şekilde çalıştırılır.

Sonuç olarak sınıflandırıcılar, herhangi bir cümlenin özete uygun olup olmayacağına karar verecektir. Bunu yapması için kullanılan öznitelikler cümlenin bulunduğu metinle alakalı olduğu için, metinden bağımsız cümle sınıflandırma olarak görülemez. Bu durumda sınıflandırıcının değerlendirmesinin yaparken metinler halinde çalıştırmak doğru sonuçlar verecektir.

Veri kümesinin 130 metin ve açıklamadan oluştuğunu varsayarsak, 5li çapraz doğrulama yöntemi ile çalışma yapabiliriz. Yani 104'ü eğitim 26sı test olmak üzere haber metni-açıklaması şeklinde veri kümesini bölüm 5 farklı çalışma yapabiliriz. Değerlendirme sonuçlarına göre sistemimize en uygun modele karar veririz.

**Referans Çalışma (Baseline)**

Oluşturulacak olan sistemimize karşı Kutlu vd. [4] nin yaptığı cümle seçerek Türkçe genel metin özetleme çalışması referans olarak alınır. Çünkü Türkçedeki sayılı metin özetleme çalışmalarında en güncel ve başarılı duyarlılık ile kesinlik sonuçları elde eden bir çalışmadır. İlgili çalışma bizimkinin aksine genel(generic) metin özeti üzerinedir. Belirli özniteliklere göre cümleler, puanlama fonksiyonu yardımıyla puanlandırılıp en yüksek puanlı cümleler özet için seçilmektedir. Öznitelikler eğitim kümesinde en yüksek puanı sağlayacak şekilde ağırlıklandırılarak puanlama fonksiyonu geliştirilmiştir. Aynı şekilde Uzundere vd. [3] nin çalışmaları da benzer puanlama yöntemi olduğu için beraber referans yöntem olarak kullanılır.

Referans çalışmada haber sitelerinden metinler özetleme için kullanılmış olup referans özetleri ise kitle kaynağı tarafından oluşturulmuştur. Bizim çalışma sürecimizde kitle kaynağı kullanılamayacağı için, kendi kullandığımız referans özetleri o çalışmada da kullanılır. Yine referans çalışma veri kümesi için kendi veri kümemizdeki metinler kullanılır. Referans çalışma değerlendirme olarak ROUGE kullanmış olup, aynı şekilde yine değerlendirilecektir.

## DENEY SONUÇLARI VE DEĞERLENDİRME

Sistemimizde haber metinlerinden cümle seçerek özetleme yapılacaktır. Bu durumda en önemli durum özetleme için en uygun cümlelerin seçilmesidir. Seçilen cümlelerin uygun olup olmadığını ise daha önceden belirttiğimiz gibi referans özetlerimiz olan haber açıklamalarına göre yapacağız. Önişlemlerde haber metin cümlelerinden kosinüs benzerliği yardımıyla haber açıklamalarına en yakın olanlar işaretlenerek ikili sınıflandırma yapıldı. Sistemimizde oluşan özetler yani seçilen cümleler de bulundukları sınıflara göre değerlendirilecek. İkili sınıflandırma yapısını kullanmak için, özetleme için uygun cümleleri seçerken SVM (Destekçi Vektör Makinası) ve RF (Rassal Orman) sınıflandırma yöntemleri kullanmaktayız. Sistemimiz sınıflandırıcılardan oluştuğu için, sınıflandırıcıları tarafından kullanılan özniteliklerin ve sınıflandırıcı parametrelerin sistemimizi nasıl etkilediği üzerine deneyler yapılarak değerlendirmelerde bulunuldu.

**Veri Kümesi**

Türkçe haber sitelerinden Scrapy[18] kullanılarak veri kümemiz oluşturuldu. NTV, Milliyet, Hürriyet, Independent Türkçe ve CNN Türk gibi Türkçe haber yayımlayan haber sitelerinden özetleme sistemimize uygun olacak şekilde haber açıklamasına sahip haber metinlerini edindik. Veri kümesi oluşturulurken herhangi bir konu kısıtlaması olmadan öncelikle yakın zamanlı haberler olmak üzere metin uzunluklarına bakılmaksızın seçildi.

Veri kümesi detayları Tablo 1'de gösterilmiştir. Ayrıca veri kümemiz, referans haber özetleme veri kümesi olarak bu çalışmamızın ana katkılarından birini oluşturmaktadır.

| Varlık | Değer |
|---|---|
| Haber Sayısı (Metin ve Açıklama) | 130 |
| Toplam Haber Açıklama Cümle Sayısı | 191 |
| Toplam Haber Metni Cümle Sayısı | 1556 |
| Toplam Kelime Sayısı | 23521 |

Tablo 1

**Deney Sonuçları ve Değerlendirme**

Geliştirdiğimiz sistemde, cümleler özet için uygun veya değil olarak ikili şekilde seçilecektir. Biz çalışmamızda, sistemimizin ürettiği özetler yani seçtiği cümleler S ile referans özetleri R karşılaştırmak için duyarlılık (1) ve kesinlik (2) ölçütlerini kullandık.

Veri kümemizi 5 parçaya ayırdık. Her çalışmada 4 parçayı eğitim, 1 parçayı da test için kullanarak çapraz doğrulama yöntemini uygulayarak değerlendirmeyi gerçekleştirdik. Değerlendirme yaparken ROUGE yöntemini kullandık. ROUGE yöntemi basitçe, deney düzeneği tarafından edinilen metin ile referans özet metni arasında Jaccard benzerliği kullanarak değerlendirmede bulunan sistemdir. Tüm deney sonuçları bu yöntemle edinilmiştir.

Eğitim sürecinde aynı zamanda seçtiğimiz özniteliklerin (konum, anlatım ifadeleri, kelime sıklığı, varlık isimleri ve başlık benzerliği) çalışmadaki etkileri üzerine deneyler yaptık. Bunun için diğer deney sonuçlarımızda SVM (gamma=1, C=1) sınıflandırıcısını kullandık. Özniteliklerin sınıflandırıcıya etkisini gözlemlemek için önce tüm öznitelikleri sonra teker teker birisi olmadan deneyler yaptık, Tablo 2'de detaylandırılmıştır. Öncelikle cümle konumunun en önemli öznitelik olduğu gözlenmektedir. Haber metinlerinde genellikle ilk cümleler özet niteliğinde olabildiği için önemli olmaktadır. Anlatım ifadeleri öznitelik olmaktan çıkartıldığında kesinlik ve duyarlılık ifadelerinin arttığı gözlemlenmektedir. Bu durumda, anlatım öznitelik değerlerinin edilmesinin yanlış fikir olduğu sonucuna varılabilir, belki bu ifadeler haber metninde fazla yer kapladığı için bu durum gerçekleşmiştir. Kelime sıklığı olmadığında, duyarlılık ve kesinlikte çok az bir azalma olmuştur. Yani terim sıklığının az da olsa özet cümlesi seçmede önemi vardır, haber metinlerindeki kelime çeşitliğinden dolayı az olmuş olabilir. Varlık isimleri de etkisi az olan özniteliktendir, bunun sebebi haber metinlerin her yerinde varlık isimlerinin geçmesi olabilir. Kesinlikte ve duyarlılıkta biraz azalma olmaktadır.

$$1)\ Duyarlılık = \frac{S \cap R}{R}$$

Çalışmalar sırasında scikit [2] ile geliştirilen SVM ve RF sınıflandırıcılarını kullandık. Her iki sınıflandırıcı da farklı parametreler ile çalıştırıldı. Detayları Figure-6'da verilen SVM çalışmaları incelendiğinde, kesinlik bakımından gamma 1'de %30'a yakın ideal sonuçlar edinilmektedir. Gamma 100'e yaklaştıkça en düşük değerlerin azaldığı görülmektedir, bu durumda aşırı uyum yaşanmış olabilir. Duyarlılık konusunda yine gamma-1'de ortalama %32 ile ideal değerler elde edilmiştir ve benzer şekilde gamma arttıkça kesinlik azalmaktadır. Ceza parametresinin (C) kesinlik ve duyarlılıkta ortalama değerler edindiğinden en ideal 1 olduğu görülmektedir ancak genel olarak çok az etki ettiği gözlemlenmektedir.

$$2)\ Kesinlik = \frac{S \cap R}{S}$$

| Ölçüt | Hepsi | Anlatım Olmadan | TF olmadan | Varlık İsimleri Olmadan | Konum Olmadan | Başlık Olmadan |
|---|---|---|---|---|---|---|
| Duyarlılık | 33.8 | 35.6 | 32.6 | 32.3 | 28.6 | 32.5 |
| Kesinlik | 29.8 | 30.9 | 28.0 | 28.2 | 26.0 | 27.9 |

Tablo 2

RF sınıflandırıcısı, kendi içerisinde ayrı veri kümeleri oluşturup onları rastgele seçerek eğitim ve test yaptığından, her çalıştırmada farklı ama benzer sonuçlar vermektedir. Bu yüzden bu sınıflandırıcı kullanılırken, her deney 10 defa tekrarlanarak ortalaması alındı. Parametre değişiklerinde dikkate değer farklılıklar tahminci(estimator) sayısında ve maksimum derinlikte Figure 7'deki gibi gözlemlenmektedir. Tahminci(ağaç) sayısı arttığında öğrenme de arttığından dolayı kesinlik değerinde çok az bir artış olmuştur. Derinlik ise arttıkça kesinlik hafif azalmıştır. Nedeni ise, derinliğin arttıkça yeni verileri öğrenmesinin zorlaşması olabilir. Duyarlılıkta ise ağaç sayısı ve derinlik pek etkili olmamıştır.

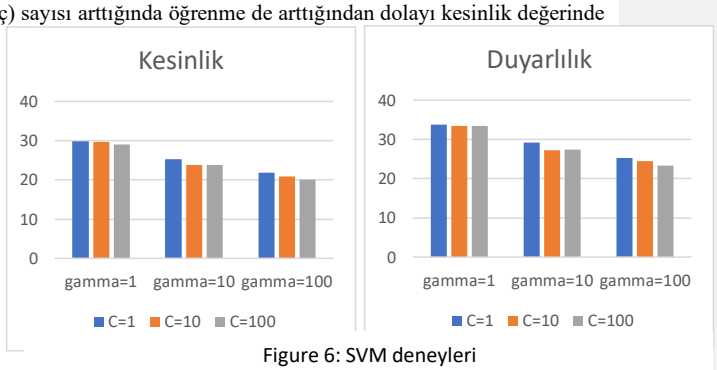


Figure 6: SVM deneyleri

Türkçe haber özetleme ve metin özetlemedeki diğer referans metotlar [3][4] ile değerlendirme yaptık. İkisi de benzer özniteliklerle cümleleri puanlayıp en yüksek puana sahip olanları özetleme için seçmektedir. Bu yüzden biz de iki yöntemi beraber kullanarak değerlendirme yaptık. Farklılık olarak Kutlu vd. çalışmalarında, öznitelikleri ağırlıklandırırken eğitim kümesi üzerinde en ideal sonuçları veren ağırlıkları bulurken, Uzundere vd. sezgisel olarak ağırlıklar vermişlerdir. Biz de benzer şekilde konum, varlık ismi, merkezilik(metin ile cümle benzerliği), başlık ile benzerlik ve kelime sıklığı özniteliklerini eğitim kümesinde en ideal sonuçlar verecek şekilde en ideal ağırlıkları elde ettik. Değerlendirme yöntemi olarak referans çalışmalarda da kullanılan ROUGE yöntemi kullanılmıştır. Tablo 3'de detayları verildiği gibi, puanlama yöntemi genel olarak geliştirdiğimiz sistem ile benzer sonuçları edinmiştir. SVM modelleri yaklaşık %20 daha başarısız olsa da en iyi sonuçlarımız Rassal Ormanlar ile elde edilmiştir.

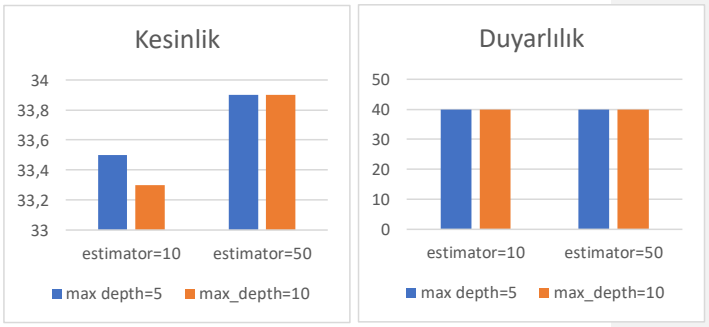


Figure 7: Rassal Orman deneyleri

| | SVM | RF | Puanlama |
|---|---|---|---|
| Duyarlılık | 32.3 | 40.3 | 39.8 |
| Kesinlik | 28.2 | 34.3 | 34.2 |
| F Ölçüsü | 28.2 | 34.7 | 34.5 |

Tablo 3

## SONUÇ

Geliştirilen sistem, Türkçe haber metinlerinden cümle seçerek özet oluşturmayı amaçlamaktadır. Bunun için sınıflandırma yöntemleri kullanılmıştır ve sınıflandırma için 5 öznitelik seçilmiştir. Sınıflandırma için etiketleme yaparken haber açıklamaları referans özet olarak kullanılmıştır. Toplam 130 haber metni ve açıklaması üzerinde çalışma yapılmış olup %40'a varan duyarlılık ve %34 kesinlik ile bu alanda yapılan diğer başarılı çalışmalara benzer sonuçlar edinilmiştir.

Sistemimiz diğer çalışmalardan farklı olarak daha büyük veri kümesi üzerinde çalışmıştır. Benzer öznitelikleri kullansa da farklı olarak sınıflandırma yöntemlerini kullanmıştır. Bu bakımdan puanlama için ağırlıkların bulunmasıyla uğraşmak yerine doğrudan sınıflandırıcılar kullanmak geliştirme açısından kolaylık sağlamıştır.

Gelecek çalışmalarda, daha büyük veri kümesi üzerinde çalışmak planlanmaktadır. Her ne kadar diğer çalışmalardan fazla verimiz olsa da sınıflandırma yöntemlerimiz için çok daha fazla veri gerekmektedir. Aynı zamanda daha detaylı çalışma ile özniteliklerin daha iyi incelenmesi ve ön işlemlerin daha detaylı yapılması gerektiğini düşünmekteyiz, bu sayede sistemimiz en ideal hale getirilebilir.

## REFERANSLAR


[1] (2018). *10 Yılda Ne Değişti?*. https://interaktif.konda.com.tr/tr/HayatTarzlari2018/#firstPage sayfasından edinildi.

[2] Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out.

[3] Uzundere, E., Dedja, E., Diri, B., & Amasyalı, M. F. (2008). Türkçe haber metinleri için otomatik özetleme. Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu, ASYU.

[4] Kutlu, M., Cığır, C., & Cicekli, I. (2010). Generic text summarization for Turkish. The Computer Journal, 53(8), 1315-1323.

[5] Keneshloo, Y., Ramakrishnan, N., & Reddy, C. K. (2019, May). Deep Transfer Reinforcement Learning for Text Summarization. In Proceedings of the 2019 SIAM International Conference on Data Mining (pp. 675-683). Society for Industrial and Applied Mathematics.

[6] Dutta, M., Das, A. K., Mallick, C., Sarkar, A., & Das, A. K. (2019). A Graph Based Approach on Extractive Summarization. In Emerging Technologies in Data Mining and Information Security (pp. 179-187). Springer, Singapore.

[7] Khan, R., Qian, Y., & Naeem, S. (2019). Extractive based Text Summarization Using K-Means and TF-IDF.

[8] Alguliyev, R. M., Aliguliyev, R. M., Isazade, N. R., Abdi, A., & Idris, N. (2019). COSUM: Text summarization based on clustering and optimization. Expert Systems, 36(1), e12340.

[9] Mandal, S., Singh, G. K., & Pal, A. (2019). PSO-Based Text Summarization Approach Using Sentiment Analysis. In Computing, Communication and Signal Processing (pp. 845-854). Springer, Singapore.

[10] Yadav, S., & Shah, D. (2018). News Summarization using Text Mining.

[11] Abujar, S., Hasan, M., & Hossain, S. A. (2019). Sentence Similarity Estimation for Text Summarization Using Deep Learning. In Proceedings of the 2nd International Conference on Data Engineering and Communication Technology (pp. 155-164). Springer, Singapore.

[12] İzmir'de makilik alanda yangın. https://www.ntv.com.tr/turkiye/izmirde-makilik-alanda-yangin-yerlesim-yerlerini-tehdit-ediyor,0-EBZ8-DhEueTO1JTj0O_A

[13] https://github.com/ahmetax/trstop/blob/master/dosyalar/turkce-stop-words

[14] Snowball. https://snowballstem.org/

[16] Text Similarities : Estimate the degree of similarity between two texts. https://medium.com/@adriensieg/text-similarities-da019229c894

[17] Scikit-learn. https://scikit-learn.org/stable/