T.C.
İSTANBUL MEDENİYET ÜNİVERSİTESİ
MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ

MEZUNİYET ÇALIŞMASI

METİN SINIFLANDIRMASINDA DESTEK VEKTÖR MAKİNESİNİN ETKİLİLİĞİNİN YAPAY SİNİR AĞLARI İLE KARŞILAŞTIRILMASI

Hilal EMİN - 18120808003

MATEMATİK BÖLÜMÜ

Danışman

Doç. Dr. Betül HİÇDURMAZ

Temmuz, 2023
İSTANBUL

ÖNSÖZ

Bu çalışmada Metin Madenciliği ve Makine Öğrenmesi Algoritmaları kullanılarak internet üzerinden çekilen haber metinlerinin sınıflandırılması amaçlanmaktadır.

Metin madenciliğinin gerekli aşamaları uygulanarak veri seti temizlenmiştir.Veri setinin temizlenme aşamaları ;
Noktalama işaretlerinin silinmesi, sayıların silinmesi, etkisiz kelimelerin silinmesi, az geçen kelimelerin silinmesi, kelime köklerinin bulunması, kelimelerin küçük harflere çevrilmesi olarak sıralanmaktadır.

Veri seti temizlenerek makine çğrenmesi algoritmalarına hazır hale getirilmiştir. Makine öğrenmesinin denetimli makine öğrenmesi-sınıflanadırma algoritmalarından Destek Vektör Makineleri (DVM) kullanılması hedeflenmiştir. Ayrıca diğer sınıflandırma algoritmalarından Yapay Sinir Ağları da kullanılarak DVM ile aralarındaki fark incelenip, algoritma karşılaştırılması yapılmaktadır.

İçindekiler

1 GİRİŞ

Yaşadığımız bilgi ve teknoloji çağında, internet üzerinden yapılan paylaşımların artması ve büyük veri setlerinin oluşması nedeniyle aranan veriye doğru bir şekilde ve kısa sürede erişim oldukça önemli bir konu haline gelmiştir. İstenen bilgiye ait veri, görüntü, video, ses ya da metin şeklinde depolanabilmektedir. Bu sonsuz bilgi havuzunda amaçlanan şekilde veriye ulaşım için literatürde veri türüne göre kullanılan farklı algoritma ve yöntemler bulunmaktadır [1].

Metin formunda olan veri setleri üzerinde analiz işlemini gerçekleştirebilmek için Veri Madenciliğinin alt alanı olan Metin Madenciliği (MM) alanındaki teknik ve yöntemler kullanılmaktadır. MM genellikle yapısal halde olmayan metin verilerinden ilgi çekici bilgi ve anlam çıkarma işlemi olarak tanımlanır. Günümüzde farklı veri kaynakları üzerinde (haberler, sosyal medya, çevrimiçi kütüphaneler vb.), farklı MM yöntemleri kullanılarak birçok bilimsel çalışma yapılmıştır [2].

1.1 Bu Alanda Yapılmış Bazı Çalışmalar

Toraman vd. çalışmalarında, Türk haber portallarında kullanılmak üzere otomatik metin kategorizasyonu ile yüksek doğruluk derecesinde bir sınıflandırma aracı sunmayı amaçlamışlardır [3]. Bilkent Haber Portalı kullanılarak oluşturulan, farklı özelliklere sahip iki Türkçe test veri setine C4.5, KNN, Naive Bayes ve SVM yöntemleri uygulanarak sonuçları tartışılmıştır. Dört farklı yöntemin sonuçlarının karşılaştırıldığı çalışmada, haber metinlerinin sınıflandırılmasında diğer kök belirleme algoritmalarının da değerlendirilmesi önerilmiştir.

Türkçe dilbilgisi özelliklerini kullanarak web tabanlı haber metinlerinin sınıflandırıldığı çalışmada, sınıflandırıcıda kullanılan özellik vektörünün boyutu ile sınıflandırıcı başarısı arasındaki ilişki irdelenmiş ve boyut azaltılmasına rağmen başarı değerinin düşmediği bir yöntem önerilmiştir [4]. Çalışma sırasında Naive Bayes, SVM, C4.5 ve Rastgele Orman sınıflandırma metotları analiz edilerek, azaltılmış özellik vektörü kullanımında en yüksek başarı oranının Naive Bayes algoritması ile sağlandığı ifade edilmiştir.

2013 yılında, Li YM, ve arkadaşı [5] dilin sahip olduğu özelliklere dayalı öznitelik çıkarımı, TF-IDF terim puanlama, destek vektör makineleri yöntemlerini bir arada kullanarak sosyal medya veri seti üzerinde sınıflandırma gerçekleştirmişlerdir. Yaptıkları çalışma sonucunda

%90,40'lık sınıflandırma başarımı elde etmişlerdir.

Camilleri vd., metin madenciliği ile çevrimiçi haberleri analiz ederek, depremlere ilişkin içerik oluşturmayı hedeflemişlerdir [6]. Çevrimiçi haberler ile dünya çapında meydana gelen sismik olaylar arasındaki ilişkinin gerçek zamanlı olarak araştırıldığı çalışmada, deprem ile ilgili raporlardan bilgiler metin madenciliği araçları ile otomatik olarak toplanarak tanımlanmakta ve sınıflandırılmaktadır. Çalışmada, dünyanın farklı yerlerinde bulunan 23 haber ajansı tarafından yayınlanan 268.182 haber ve bültende listelenen büyüklükleri 4 ile 8.2 arasında değişen 14.717 deprem verileri kullanıldığı belirtilmiştir.

2 VERİ ÖN İŞLEME VE METODOLOJİ

2.1 Veri Seti Hakkında

Kullandığımız veri seti Türkçe haber başlıkları ve kategorilerden oluşmaktadır. Toplam 4200 adet haber başlığı ve kategorisi bulunmaktadır. Haber başlıklarına ait kategoriler ise şu şekildedir; "Ekonomi, Siyaset, Yaşam, Teknoloji, Magazin, Sağlık, Spor" olmak üzere 7 adet konu başlığı vardır. [7]

ETIKET	
Ekonomi	600
Magazin	600
Sağlık	600
Siyaset	600
Spor	600
Teknoloji	600
Yaşam	600

Şekil 1: Etiketler ve Toplam Haber Sayısı

Her bir kategori için 600 adet haber başlığı vardır.

2.2 Veri Ön İşleme (Feature Engineering)

Sınıflandırma algoritmalarının çalıştırılabilmesi için veri ile ilgili bazı veri temizleme ve düzenleme işlemlerinin gerçekleştirilmesi gerekmektedir. Yapılan işlemler genellikle metin tabanlı olmayan verileri metin içerisinden çıkartma (noktalama işaretleri, boşluk, özel karakterler vb.), küçük harfe dönüştürme, etkisiz kelimeleri (stop words) ayıklama gibi kısımlardan oluşmaktadır.

2.2.1 Veri Setinin Temizlenmesi

İlk olarak veriye ön temizlenme işlemini yapalım.

1. Noktalama İşaretlerinin Silinmesi: İlk adım noktalama işaretlerini sileriz. Çünkü noktalama işaretleri sınıflandırma sırasında bir anlam ifade etmemektedir.

2. Sayıların Silinmesi: Haber metinleri içerisinde bulunan sayıları da sileriz. Metin bazlı çalışmak istediğimiz için sayılar işimize yaramamaktadır.

3. Etkisiz Kelimelerin Silinmesi: Etkisiz kelimleri çıkarırız. Etkisiz kelimeler (stopwords);bilgi çıkarımı ve sınıflandırma için bir anlam ifade etmeyen, örneğin; Türkçe dilindeki "ve", "veya", "bazı", "hepsi" gibi kelimelerden oluşmaktadır. Diğer diller için de ayrıca etkisiz kelime listeleri oluşturulmaktadır. Veri setimiz için Python'ın Doğal Dil İşleme Kütüphanesi olan NLTK'dan corpus ile Türkçe stopwords'leri yani etkisiz kelimeleri alıp haber metinlerinden çıkardık.

4. Az Geçen Kelimelerin Silinmesi: Haberler içinde az geçen kelimeleri sileriz. Bunun nedeni veri setimizde o kadar fazla kelime olur ki, modele bunların hepsini sokmak istemeyiz. Bu yüzden az geçen kelimeleri atmamız iyi olur. Bazen de bir kelime, bir ifade o kadar çok geçer ki, kullanılan yöntemlere göre, analizi (duygu analizi gibi) çok etkileyebilir.
İlk olarak kelimelerin frekanslarını bulduk. Ardından silmek istediğimiz kadarını seçtik ve seçtiklerimizi veri setimizden sildik.

5. Kelime Köklerinin Bulunması: Metin analizi çalışmalarında özellikle metin içerisindeki her bir kelime bir terim olarak ele alındığında öznitelik sayısı metin içerisinde geçen tekil kelime sayısına denk gelmektedir. Bu durum öznitelik miktarının fazla olmasına neden olmaktadır. Bunun önüne geçebilmek için yani öznitelik sayısını düşürebilmek için terimlerin kökleri saptanmaktadır. Böylece aynı kökten türemiş kelimeler tek bir öznitelik olarak değerlendirilmekte ve neticesinde öznitelik sayısı azalmaktadır. Kök bulma işlemi bir doğal dil

işleme çalışmasıdır ve Türkçe metinler Zeyrek-NLP kullanılarak gerçekleştirilmektedir [8]. Burada öncelikle her bir haber metnini split() metotu ile kelimelere ayırdık. Ardından split() içindeki her bir elemanın köklerini bulduk ve bunları tekrar cümle haline getirdik.

6. Kelimelerin Küçük Harflere Çevrilmesi: Türkçe'de kelimenin büyük veya küçük harfli olması anlamı değiştirmediğinden bütün harfleri küçük harfe çevirириz.

Tüm bu ön işleme basamaklarını veri setimize uyguladık ve veri setinin son hali Şekil 2'deki gibi olur.

	HABERLER
0	tüik veri göre sanayi ciro endeks ağustos ayın...
1	piyasa gün ek rezerv başlamak
2	citigroup deutsche bank hsbc libor manipülâsyo...
3	gelişmek piyasa yatırım fazla fed ilgi
4	bitcoin fiyat yükselmek hız kesmek
...	...
4195	osmaniye kadir ilçe cadde yürümek tipi hoş git...
4196	ihraç edi görev el çektiri öğretmen yönelik op...
4197	konak ilçe operasyon gram esrar uyuşturucu hap...
4198	siirt manisa düzmek operasyon gözaltı alınan e...
4199	deniz kaçak sigara operasyon gözaltı

Şekil 2: Temizlenmiş Veri Seti

2.2.2 Word Cloud (Kelime Bulutu)

Metin madenciliğinde en sık kullanılan veri görselleştirme yöntemidir. Kelimelerin geçme frekanslarına göre bir şekil (bulut gibi) içinde sunulmasıdır.

2.2.3 Metin Vektörleştirme

Metin verilerini sayısal verilere/vektöre dönüştürme işlemine vektörleştirme (vectorization) veya NLP dünyasında kelime gömme (word embedding) deni. Kelime vektörleştirme yöntemleri, makine ve derin öğrenme süreçlerinden önce uygulanan bir dönüşüm işlemidir. Bu dönüşüm ile birlikte metin farklı yöntemlere göre kendi sınıfı içerisinde sayısal veriler içeren vektörlere dönüştürülür. Böylece metin verisi algoritmalar tarafından eğitilebilecek ve

analiz edilebilecek sayısal değerlere dönüştürülmüş olur.
Bu veri setinde Bag of Words (Kelime Çantası-Count Vectorizer) ve TF-IDF kullandık.

Bag of Words (Kelime Çantası-Count Vectorizer): BoW modeli, bir dokümandaki terimlerin oluşum şeklini (Örneğin: terim sayılarını) belirten metnin temsil biçimidir. Bu modelde; terim pozisyonu ve sözcük sıralaması dikkate alınmaz [9]. Veri içerisindeki tüm benzersiz kelimelerin sayısı ve verinin toplam sınıf sayısı olacak şekilde bir matris oluşturur. Hangi sınıfta o kelime var ise matristeki ilgili hücreye 1 değeri, olmayan durumda ise 0 değeri atanır.

Şekil 3: Bag Of Words

Şekil 3'te görüldüğü üzere eğer kelime cümle içerisnde geçiyorsa 1 değerini geçmiyor ise 0 değerini almaktadır.

BoW yaklaşımında terim-sayma amacıyla; bir metin dokümanı koleksiyonunu terim sayısı matrisine dönüştüren CountVectorizer sınıfını kullanacağız [10].

```
BoW_Vector = CountVectorizer(min_df = 0., max_df = 1.)
BoW_Matrix = BoW_Vector.fit_transform(df1['HABERLER'])
print(BoW_Matrix)
```

Şekil 4: Count Vectorizer

Burada ilk olarak CountVectorizer sınıfını kullanabilmek için BoW Vector adlı vektöre atadık. Ardından bunu haber metinlerine fit ve transform yaptık. BoW Matrix'i elde etmiş olduk.

```
(0, 2015)       1
(0, 622)        1
(0, 565)        2
(0, 980)        1
(0, 7583)       1
(0, 7282)       1
(0, 548)        1
(0, 411)        1
(1, 5192)       1
(1, 2784)       1
(1, 1942)       1
```

Şekil 5: BoW Matrix

Burada BoW Matrix'in bir kısmını aldık. Bu matrisi şu şekilde yorumlarız: 0.indeksteki haber metininde 2015 indeksli kelime 1 defa geçerken 565 indeksli kelime 2 defa geçmektedir.

Bulduğumuz bu diziyi toarray() ile bir matrise dönüştürdük.

	ab	abazi	abd	abdulfettah	abdulhamit	abdulkadir	abdullah	abi	abid	abla	...	şuur	şâhıs	şçil	şöhret	şöyle	şükür	şüphe	şık	şıkel	şırnak
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4195	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4196	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0
4197	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4198	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4199	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Şekil 6: Count Vectorizer Matrisi

Görüldüğü üzere veri setindeki özniteliklerimiz (kelimeler) her bir haber başlığında kaç defa geçtiyse 0'dan büyük olacak şekilde değeri yazılır eğer hiç geçmediyse 0 değeri yazılır.

TF-IDF: Terim sıklıklarının sayılması ile ilgili en önemli sorun, sık kullanılan terimlerin dokümanda baskın olmaları ve artık dokümanı temsil eder hale gelmeleridir. Bu problemi çözmek için, "Terim Frekansı x Ters Belge Frekansı" anlamına gelen "TF x IDF" modelini ve skorlama yöntemini kullanabiliriz. Hesaplamada iki ölçüt kullanır: terim sıklığı (TF) ve ters belge sıklığı (IDF). TF ve IDF hesaplama formülleri şu şekildedir;

$$\text{TF}_{(d,m)} = \frac{\text{m kelimesinin d dökümanında geçme sayısı}}{\text{dökümandaki toplam kelime sayısı}}$$

$$\text{IDF}_{(d,m)} = ln\frac{\text{toplam döküman sayısı}}{\text{içerisinde m kelimesini bulunduran toplam döküman sayısı}}$$

$$\text{TF x IDF Skoru} = TF * IDF$$

ile bulunur. Tek veya küçük bir belge grubunda ortak olan kelimeler genel kelimelerden daha yüksek TF-IDF skoruna sahip olma eğilimindedir.

TF-IDF yaklaşımında TfidfVectorizer sınıfını kullanacağız.TF-IDF'ide 3 farklı alt grup ile hesaplayacağız.

İlk önce Word Level'ı hesaplayalım. Word Level yukarıda yazdığımız TF-IDF skor formülü ile direkt hesaplanan yöntemdir.

```
Tfidf_Vector = TfidfVectorizer(min_df = 0., max_df = 1., use_idf = True)
Tfidf_Matrix = Tfidf_Vector.fit_transform(df1['HABERLER'])
print(Tfidf_Matrix)
```

Şekil 7: Word Level

TfidfVectorizer sınıfını kullanabilmek için Tfidf Vector isimli nesneye atadık. Ardından bunu haber metinlerine fit ve transform ettik. Böylece matrisi elde etmiş olduk.

```
(0, 411)      0.1916288355955901
(0, 548)      0.24485518874610004
(0, 7282)     0.1649933882304094
(0, 7583)     0.19391034943501698
(0, 980)      0.11806753574126104
(0, 565)      0.43498387739627653
```

Şekil 8: Word Level

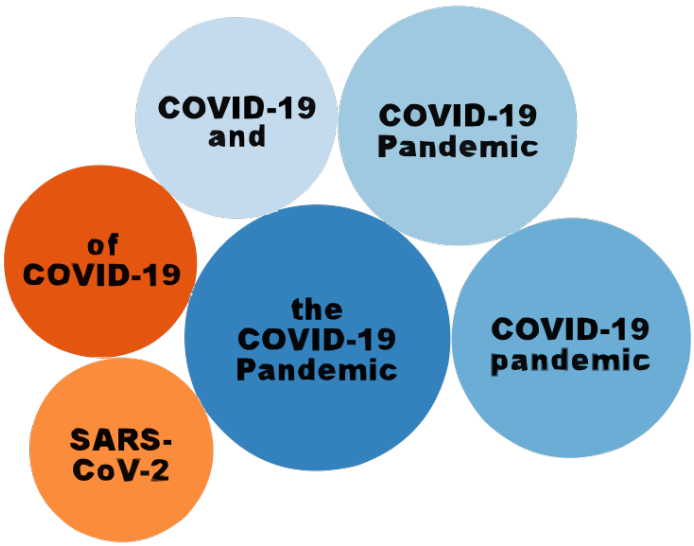
Burada Word Level Matrix'in bir kısmını aldık. Bu matrisi şu şekilde yorumlarız: 0.indeksteki haber metinindeki 411 indeksli kelime yaklaşık olarak 0.1916 TF-IDF skoruna sahiptir.

	ab	abazi	abd	abdulfettah	abdulhamit	abdulkadir	abdullah	abi	abid	abla	...	şuur	şâhıs	şçil	şöhret	şöyle	şükür	şüphe	şık	şıkel	şırnak
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4195	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.0	0.0	0.0
4196	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.216	0.0	0.0	0.0
4197	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.0	0.0	0.0
4198	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.0	0.0	0.0
4199	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.0	0.0	0.0

Şekil 9: Word Level Matrisi

Tüm kelimeler için TF-IDF skorlarını gösteren matris.

İkinci olarak N-Gram TF-IDF uygulayalım. N-Gram; n adet elemandan oluşan ardışık dizilere verilen genel addır. Doğal dil işleme ve hesaplamalı dilbilim bağlamında ngramları oluşturan elemanlar, ihtiyaca ve uygulama alanına göre bir konuşma metni ya da yazılı metin içindeki kelimeler, heceler, ses birimleri ya da harfler olarak seçilebilir.



Şekil 10: Coronavirus Hastalığı N-Gram

Şekil 10'a bakalım. Burada Coranavirüs hastalığıyla ilgili yayınlarda Coranavirüs ile beraber sıklıkla kullanılan 6 farklı kelime verilmiştir.
1 boyutundaki bir n-gram "unigram" olarak adlandırılır; boyut 2 bir "bigram"dır (veya daha az yaygın olarak bir "digram"); boyut 3 bir "trigram" dır.
Veri setimize bigram ve trigram uygulayalım. TfidfVectorizer()'ın içine ngram aralıklarımız olan (2,3)'ü girdik ve bunların TF-IDF skorlarınu hesapladık. Ardından bunu bir matrise dönüştürdük.

	ab bakmak	ab bakmak baş	ab bakmak gıda	ab bakmak çelik	ab bakmak ömer	ab bütçe	ab bütçe türkiye	ab komisyon	ab komisyon başka	ab konsey	...	şırnak cehennem	şırnak cehennem dermek	şırnak cizre	şırnak cizre ilçe	şırnak kır	şırnak kır saldırı	ş
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
4195	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
4196	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
4197	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
4198	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
4199	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	

Şekil 11: N-Gram'ın Bir Kısmı

Buradan matrisi şu şekilde yorumlayabiliriz; "ab" ile "bakmak" kelimesinin TF-IDF skoru hesaplanmış. Ardından "ab","bakmak" ve "baş" bu 3 kelimenin beraber olduğu TF-IDF skorları hesaplanmış. Bu şekilde beraber sık kullanılan kelimelerin TF-IDF skorlarını hesapladık.

Üçüncü olarak Characters bazlı N-Gram TF-IDF uygulayalım. Characters N-Gram; n-gramlar heceler ya da harfler olarak seçilir.

	a	a	ab	ac	ad	ae	af	ag	ah	ai	...	ze	zl	zm	zn	ç	çe	çi	ş	şb	şç
0	0.148	0.0	0.0	0.0	0.000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.000	0.0	0.0	0.0	0.000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.000	0.0	0.0	0.0	0.000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.000	0.0	0.0	0.0	0.000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.000	0.0	0.0	0.0	0.000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4195	0.027	0.0	0.0	0.0	0.071	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4196	0.054	0.0	0.0	0.0	0.071	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4197	0.000	0.0	0.0	0.0	0.000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4198	0.037	0.0	0.0	0.0	0.000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4199	0.000	0.0	0.0	0.0	0.000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Şekil 12: Char N-Gram

Burada n-gram ile benzer şekilde yorumlama yapılır. "a" harfi ile yan yana kullanılan en sık karakterlerin TF-IDF skorları hesaplanmaktadır.

2.2.4 Veri Setinin Train ve Test Ayrımı

Kullandığımız veri setini Train ve Test olmak üzere 2 kısıma ayırırız. Train ve Test nedir önce bunlardan bahsedelim.

Train Set (Eğitim Seti): Train veri seti üzerinde temel modelleme denemeleri yapılarak en doğru makine öğrenmesi algoritması seçilmeye çalışılır. Bu veri seti üzerinden en fazla örneklem alınan gözlemlerden oluşur. Modelin ne kadar doğru öğrendiğini test etmek ve seçilen algoritma doğrultusunda daha önce görmediği veri seti girdi olara verildiğinde çıktısnın ne olduğunu en yüksek doğruluk ile tahmin etmektedir.

Test Set (Test Seti): Train veri setinden kalan bölüm Test veri setidir. Bir eğitim kümesinde geliştirilen modeli değerlendirmek için kullanılan bir veri kümesidir.
Eğitim setimiz büyüdükçe modelimiz daha iyi öğrenecektir. Test setimiz büyüdükçe ise değerlendirme metriklerimize daha iyi güvenebileceğimiz daha sıkı güven aralıklarımız olacaktır. Genellikle veri seti %70 -%80 arasında Train Set için bölünürken kalanı ise Test Set için bölünür.

Verimizde y değişkeni benim hedef değişkenim (target variable) yani tahminde (haberlerin etiketleri) bulunacağım değişken. X (haberler) ise veri kümesinin tüm özellikleridir. Kısaca x (haberler) bağımsız değişken y (etiketler) bağımlı değişkenlerdir.Verilerimizi bölmek için **train test split** fonksiyonunu kullanıyoruz. Burada kodumuzu her çalıştırdığımızda veri setimiz her seferinde %75'i train ve %25'i test olmak üzere ayrılır.

3 DESTEK VEKTÖR MAKİNELERİ İLE MODEL BELİRLEME

3.1 Destek Vektör Makinesi

DVM, Vapnik - Chervonenkis tarafından geliştirilen, sınıflandırma başta olmak üzere kümeleme ve regresyon problemlerinde kullanılan bir makine öğrenmesi modelidir. DVM modelinin amacı temel olarak, hedef değişkene ait sınıfları birbirinden en uygun şekilde ayıracak hiperdüzlemi tespit etmektir.
DVM, verinin doğrusal ya da doğrusal olmayan şekilde sınıflandırılmasına göre iki durumda incelenir [11].

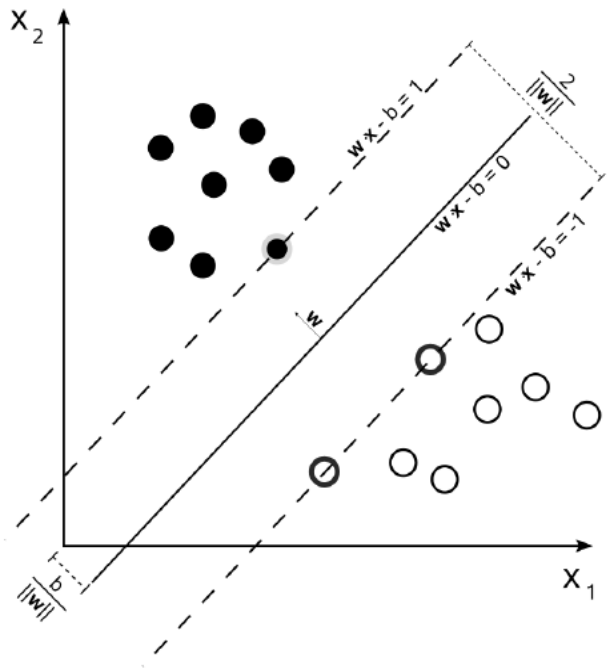
3.1.1 Doğrusal (Lineer) DVM

$(\vec{x_1}, y_1), (\vec{x_2}, y_2), ..., (\vec{x_n}, y_n)$ bir grup gözlemden oluşan bir veri kümesi olsun. y_i değerlerinin her biri -1 ya da $+1$ sayısını alması ile her bir x_i değerine karşılık gelen bir sınıfı ifade etmektedir. Her x_i değeri p boyutlu bir vektördür.

Bu iki sınıfın arasındaki maksimum uzaklığa marjin adı verilmektedir. Bu iki sınıfı ayıracak sonsuz sayıda doğru bulunmaktadır. Fakat marjini maximize eden sadece bir tane doğru bulunmaktadır. Bu doğru en iyi ayırıcı hiper düzlem olarak adlandırılmaktadır.Maksimum aralıklı hiperdüzlemi, $y_i = 1$ değeri için x_i gözlem noktalarını, $y_i = -1$ için incelenen gözlem noktalarından en uygun biçimde ayırmaktadır. Bu hiperdüzlemde, her iki grubun en yakın noktası arasındaki mesafenin en üst düzeye çıkarılması amaçlanmaktadır. Hiperdüzlemin üstünde ve altında bulunan kesikli çizgiler ise sınır düzlemleridir.Bu sınır düzlemelerin üstünde bulunan veriler ise destek vektröleri olarak adlandırılır. Hiperdüzlem bu iki sınır düzlemin ortasından geçmeli ve ikisine de eşit mesafede olmalıdır.

Herhangi bir hiperdüzlem, $\vec{w}.\vec{x} = 0$ eşitliğini sağlayan $\vec{x}$ noktalar kümesi olarak yazılabilir. Burada $\vec{w}$; hiperdüzleme doğru bir ağırlık vektörünü göstermektedir, $\frac{b}{||\vec{w}||}$ parametresi; $\vec{w}$ normal vektörü boyunca orijinden geçen hiperdüzlemi, b; eğilim değerlerini ve $\frac{2}{||\vec{w}||}$ ise; $w^T x^{(i)} + b = -1$ ile $w^T x^{(i)} + b = +1$ arası uzaklığı ifade etmektedir.

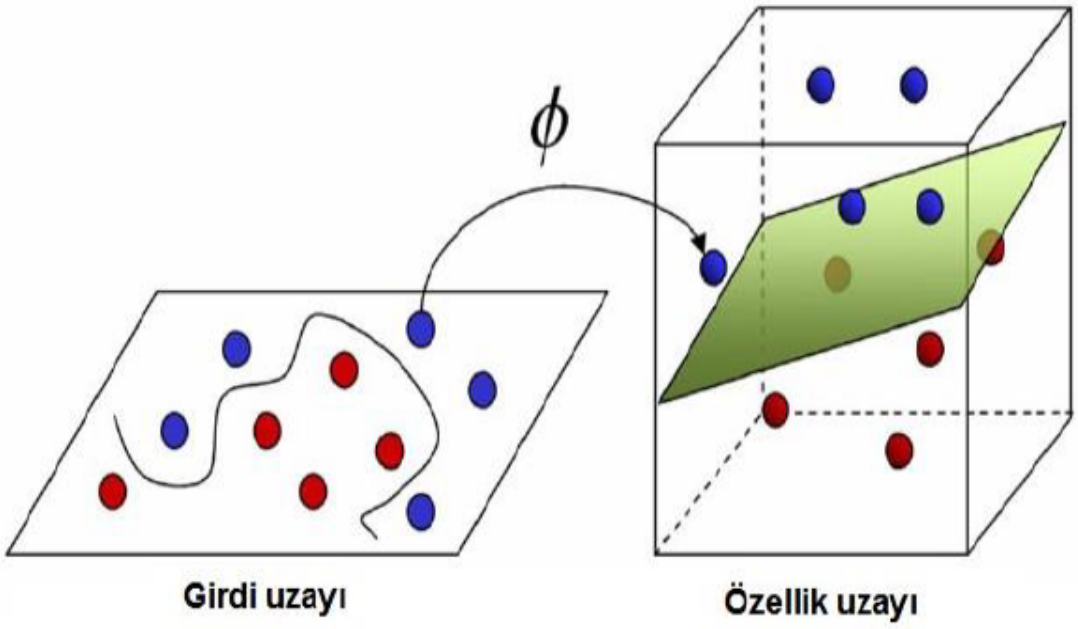
Bu yapının grafiksel gösterimi Şekil 13'te verilmiştir [11].



Şekil 13: Doğrusal DVM Modeli

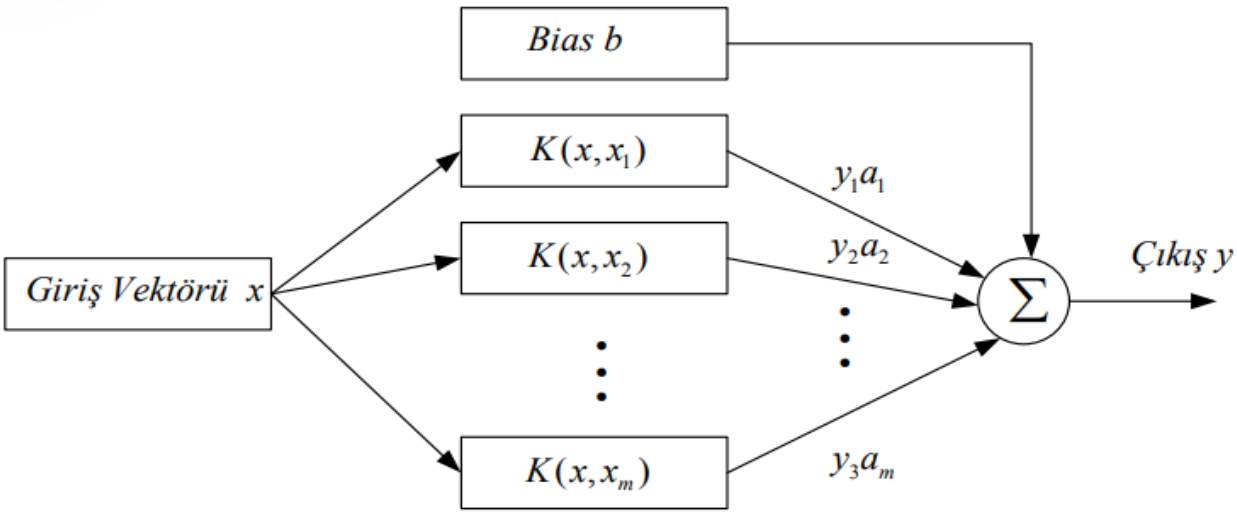
3.1.2 Doğrusal Olmayan DVM

Doğrusal olmayan DVM yönteminde, değişik yapıdaki çekirdek fonksiyonları maksimum aralıklı hiperdüzleme uygulanarak doğrusal olmayan sınıflandırıcılar elde edilmektedir. Ortaya çıkan algoritma, doğrusal DVM ile benzerdir; ancak, her iç çarpımı doğrusal olmayan bir çekirdek fonksiyonu ile değiştirilmiştir. Böylelikle algoritma, maksimum aralıklı hiperdüzlemin dönüştürülmüş örnek uzayına yerleşmesine izin vermektedir. Söz konusu dönüşüm, doğrusal yapıda olmayabilir ve dönüştürülmüş örnek uzayı ise yüksek boyutlu olabilir. İncelenen sınıflandırıcı dönüştürülmüş örnek uzayında bir hiperdüzlem olmasına rağmen, orijinal girdi uzayında doğrusal olmayabilir.Doğrusal olarak sınıflandırılamayan girdi uzayının bir üst boyuta çekirdek fonksiyonu ile haritalanarak doğrusal olarak sınıflandırılması Şekil 14'te gösterilmektedir [11].



Şekil 14: Doğrusal Olarak Sınıflandırılamayan Girdi Uzayının Bir Üst Boyuta Çekirdek Fonksiyonu İle Haritalandırılması

Bu sınıflandırmanın arkasındaki matematikten bahsedelim.



Şekil 15: DVM'in Ağ Yapısı

$K(x_i, x_m)$ çekirdek fonksiyonlarını, $y_i a_m$'ler α olarak adlandılır Lagrange çarpanlarını göstermektedir. Çekirdek fonksiyonları yardımıyla girdilerin iç çarpımları hesaplanmaktadır.

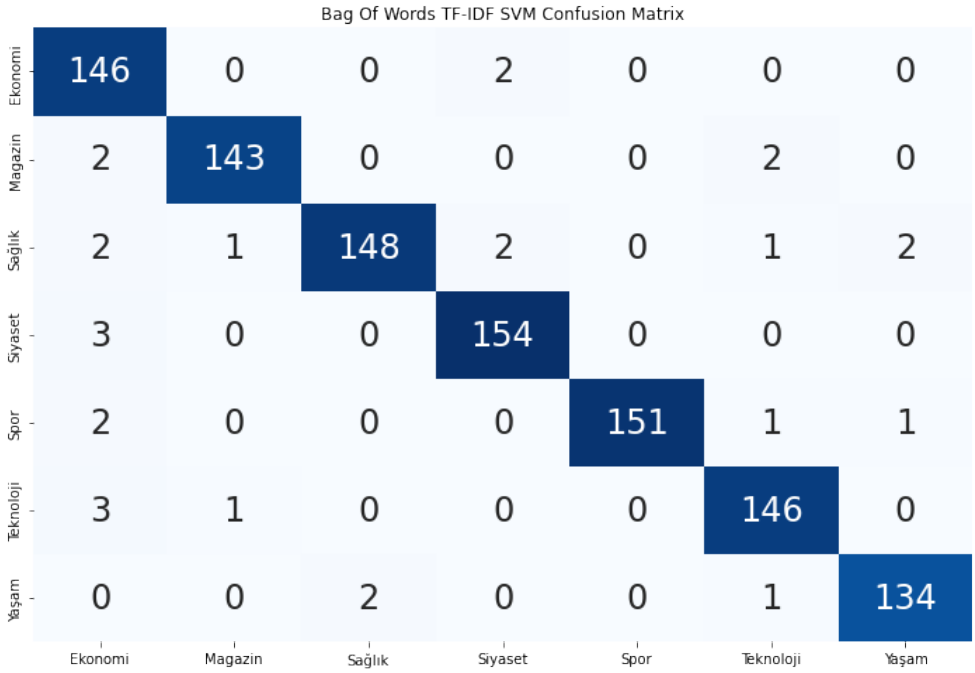
Lagrange çarpanları ise ağırlıkları göstermektedir. DVM'de bir örneğe ilişkin çıktı değeri, girdilerin iç çarpımları ile Lagrange çarpanlarının bağımsız kombinasyonlarının toplamına eşittir.

3.2 Doğrusal DVM İle Modellerin Test Edilmesi

İlk olarak modelimizi doğrusal DVM ile test ediyoruz. En sık kullanılan çekirdeklerden biridir. Belirli bir veri kümesinde çok sayıda özellik olduğunda çoğunlukla kullanılır. Birçok özelliğin bulunduğu örneklerden biri, her harf yeni bir özellik olduğu için Metin Sınıflandırma'dır. Bu yüzden Metin Sınıflandırmada çoğunlukla Doğrusal Çekirdek kullanıyoruz. Vektörleştirme işleminde kullandığımız vektörleştirme çeşitlerini tek tek DVM içerisine koyup en iyi sonuç vereni seçmeye çalışacağız.

3.2.1 DVM'in Bag Of Words İle Test Edilmesi

1) İlk olarak vektörleştirme için kullandığımız **CountVectorizer** fonksiyonunu tanımladık. Ardından train setini fit ve transform ile, test verimizi ise sadece transform ile yeni değişkene atadık.
2) DVM kullanmak için **sklearn** kütüphanesinden SVC'yi import ettik. Çekirdek olarak da lineer çekirdeği seçtik ve bunu bag of words ile vektörleştirdiğimiz train verimize ve bunların karşılık geldiği kategorilere fit ettik.
3) Ardından modelin accuracy yani doğruluk değerini hesaplamak için test için ayırdığımız verileri kullandık ve hesapladık.
4) Modelimizin doğruluk değerini, karışıklık matrisini ve sınıflandırma raporlarını bulduk.



Şekil 16: Karışıklık Matrisi

Bu karışıklık matrisinde x ekseni tahmin edilen kategorilere, y ekseni ise test verisine aittir. Bu matrisi soldan sağa yorumlarınır. Örneğin ekonomi kategorisinden test edilen toplam 148 verinin 146 tanesi doğru sınıflandırılırken 2 tanesi siyaset kategorisinde sınıflandırılmıştır. Spor kategorisinde ise toplam 155 adet test edilen veriden 151 tanesi doğru sınıflandırılırken 2 tanesi ekonomi, 1 tanesi teknoloji, 1 tanesi yaşam kategorisinde sınıflandırılmıştır.

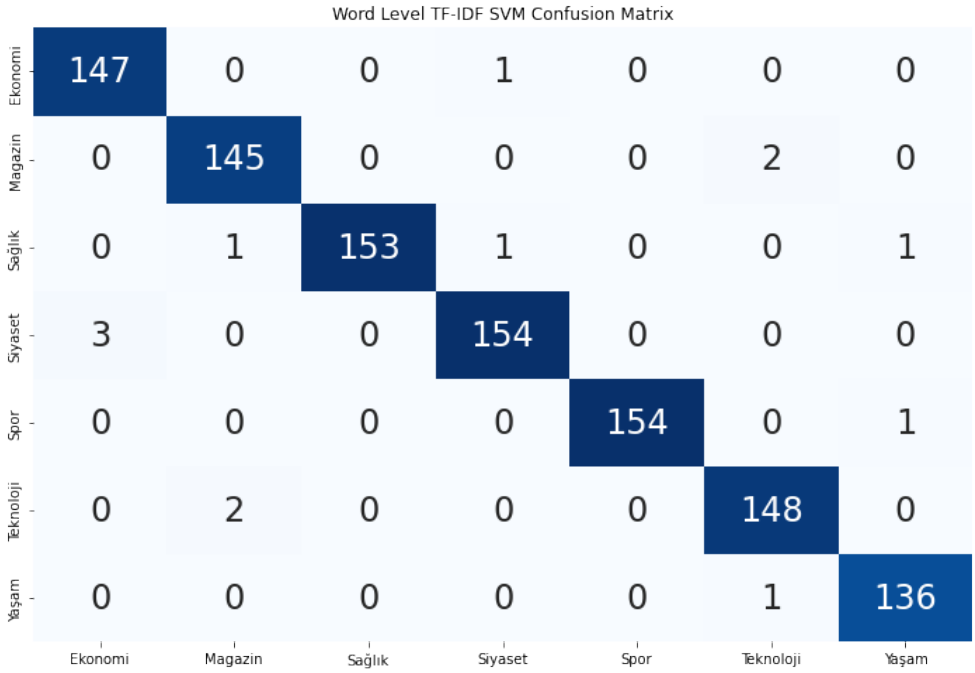
Tablo 1: Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.92	0.99	0.95	148
Magazin	0.99	0.97	0.98	147
Sağlık	0.99	0.95	0.97	156
Siyaset	0.97	0.98	0.98	157
Spor	1.00	0.97	0.99	155
Teknoloji	0.97	0.97	0.97	150
Yaşam	0.98	0.98	0.98	137
accuracy			0.97	1050
macro avg	0.97	0.97	0.97	1050
weighted avg	0.97	0.97	0.97	1050

Sınıflandırma raporundan da görüldüğü üzere accuracy yani modelin doğruluk yüzdesi %97'dir.

3.2.2 DVM'in Word Level İle Test Edilmesi

1) İlk olarak vektörleştirme için kullandığımız **TfidfVectorizer** fonksiyonunu tanımladık. Ardından train setini fit ve transform ile, test verimizi ise sadece transform ile yeni değişkene atadık.
2) DVM kullanmak için **sklearn** kütüphanesinden SVC'yi import ettik. Çekirdek olarak da lineer çekirdeği seçtik ve bunu word level ile vektörleştirdiğimiz train verimize ve bunların karşılık geldiği kategorilere fit ettik.
3) Ardından modelin accuracy yani doğruluk değerini hesaplamak için test için ayırdığımız verileri kullandık ve hesapladık.
4) Modelimizin doğruluk değerini, karışıklık matrisini ve sınıflandırma raporlarını bulduk.



Şekil 17: Karışıklık Matrisi

Bu matrise göre magazin haberlerinden test edilen 147 adet verinin 145 tanesi doğru sınıflandırılırken 2 tanesi teknoloji kategorisinde sınıflandırılmıştır. 157 adet siyaset verisinin ise 154 tanesi doğru sınıflandırılmış fakat 3 tanesi ekonomi kategorisinde sınıflandırılmıştır.

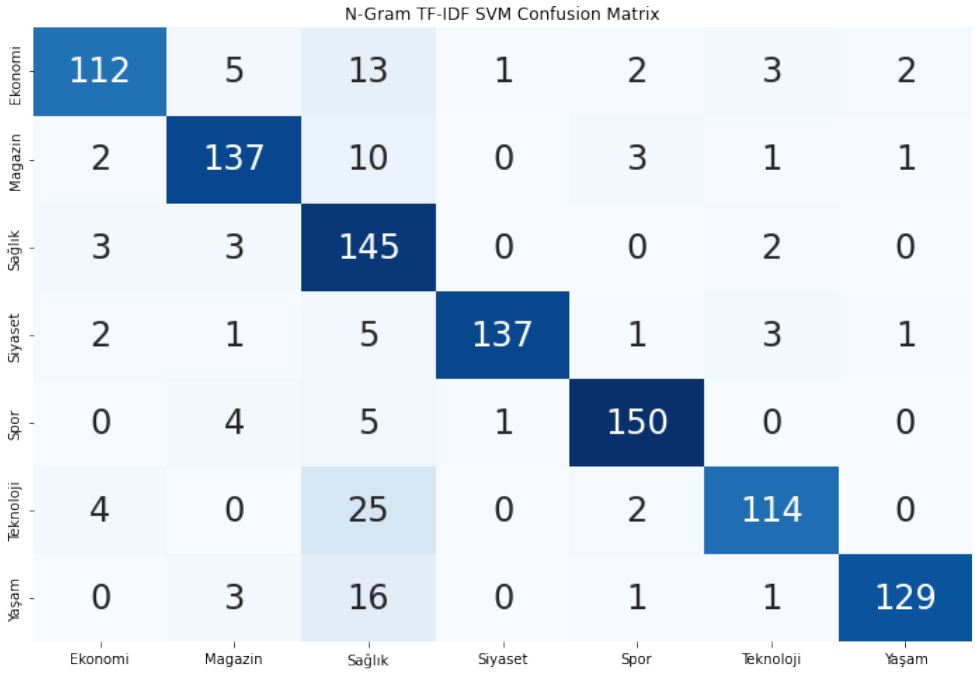
Tablo 2: Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.98	0.99	0.99	148
Magazin	0.98	0.99	0.98	147
Sağlık	1.00	0.98	0.99	156
Siyaset	0.99	0.98	0.98	157
Spor	1.00	0.99	1.00	155
Teknoloji	0.98	0.99	0.98	150
Yaşam	0.99	0.99	0.99	137
accuracy			0.98	1050
macro avg	0.99	0.99	0.98	1050
weighted avg	0.99	0.99	0.98	1050

Modelin sınıflandırma raporuna göre doğruluk değeri %98 olarak verilmektedir.

3.2.3 DVM'in N Gram İle Test Edilmesi

1) İlk olarak vektörleştirme için kullandığımız **TfidfVectorizer** fonksiyonunun ngram range vektörleştirme işlemini tanımladık. Ardından train setini fit ve transform ile, test verimizi ise sadece transform ile yeni değişkene atadık.
2) DVM kullanmak için **sklearn** kütüphanesinden SVC'yi import ettik. Çekirdek olarak da lineer çekirdeği seçtik ve bunu n gram ile vektörleştirdiğimiz train verimize ve bunların karşılık geldiği kategorilere fit ettik.
3) Ardından modelin accuracy yani doğruluk değerini hesaplamak için test için ayırdığımız verileri kullandık ve hesapladık.
4) Modelimizin doğruluk değerini, karışıklık matrisini ve sınıflandırma raporlarını bulduk.



Şekil 18: Karışıklık Matrisi

Bu matrise göre teknoloji haberlerinden test edilen 145 adet verinin ancak 114 tanesini doğru sınıflandırabilmiştir. 25 tanesi sağlık kategorisinde, 4 tanesi de ekonomi kayegorisinde sınıflandırılmıştır. 150 adet yaşam verisinin ise 129 tanesi doğru sınıflandırılmış fakat 3 tanesi magazin, 16 tanesi sağlık kategorisinde sınıflandırılmıştır.

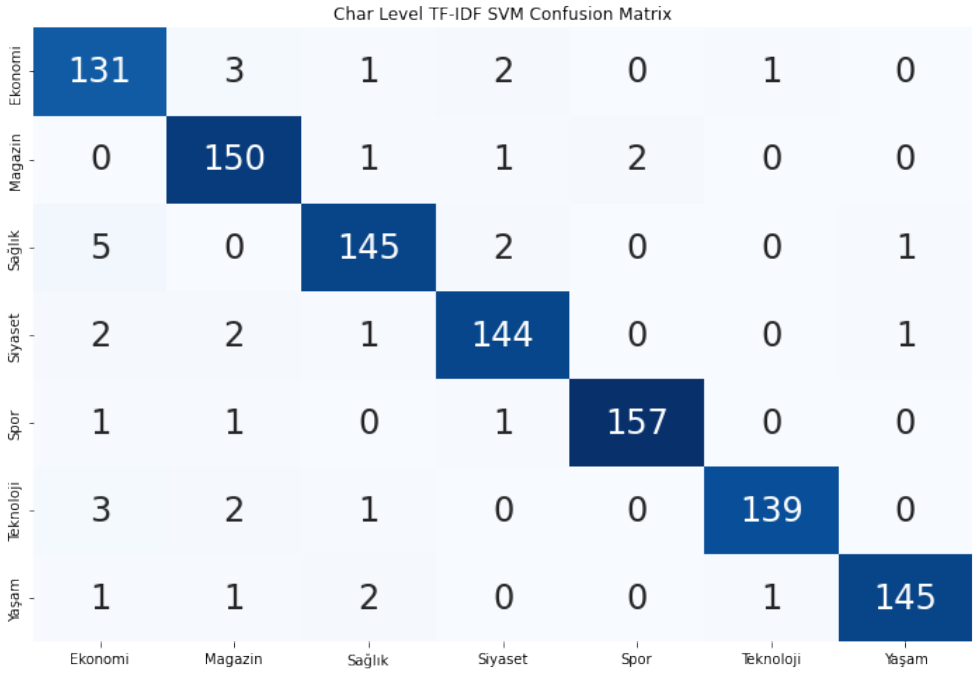
Tablo 3: Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.91	0.81	0.86	138
Magazin	0.90	0.89	0.89	154
Sağlık	0.66	0.95	0.78	153
Siyaset	0.99	0.91	0.95	150
Spor	0.94	0.94	0.94	160
Teknoloji	0.92	0.79	0.85	145
Yaşam	0.97	0.86	0.91	150
accuracy			0.88	1050
macro avg	0.90	0.88	0.88	1050
weighted avg	0.90	0.88	0.88	1050

Modelin sınıflandırma raporuna göre doğruluk değeri %88 olarak verilmektedir.

3.2.4 DVM'in Char Level İle Test Edilmesi

1) İlk olarak vektörleştirme için kullandığımız **TfidfVectorizer** fonksiyonunun analyzer='char' vektörleştirme işlemini tanımladık. Ardından train setini fit ve transform ile, test verimizi ise sadece transform ile yeni değişkene atadık.
2) DVM kullanmak için **sklearn** kütüphanesinden SVC'yi import ettik. Çekirdek olarak da lineer çekirdeği seçtik ve bunu n gram ile vektörleştirdiğimiz train verimize ve bunların karşılık geldiği kategorilere fit ettik.
3) Ardından modelin accuracy yani doğruluk değerini hesaplamak için test için ayırdığımız verileri kullandık ve hesapladık.
4) Modelimizin doğruluk değerini, karışıklık matrisini ve sınıflandırma raporlarını bulduk.



Şekil 19: Karışıklık Matrisi

Bu matrise göre siyaset haberlerinden test edilen 150 adet verinin 144 tanesini doğru sınıflandırabilmiştir. 2 tanesi ekonomi,2 tanesini magazin ve 1 tanesini sağlık kategorisinde sınıflandırılmıştır. 138 adet ekonomi verisinin ise 131 tanesi doğru sınıflandırılmış fakat 3 tanesi magazin, 1 tanesi sağlık, 2 tanesini siyaset ve 1 tanesini teknoloji kategorisinde sınıflandırılmıştır.

Tablo 4: Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.92	0.95	0.93	138
Magazin	0.94	0.97	0.96	154
Sağlık	0.96	0.95	0.95	153
Siyaset	0.96	0.96	0.96	150
Spor	0.99	0.98	0.98	160
Teknoloji	0.99	0.96	0.97	145
Yaşam	0.99	0.97	0.98	150
accuracy			0.96	1050
macro avg	0.96	0.96	0.96	1050
weighted avg	0.96	0.96	0.96	1050

Modelin sınıflandırma raporuna göre doğruluk değeri %96 olarak verilmektedir.

Eğer lineer çekirdek fonksiyonu için bir sıralama yaparsak en iyi performans değeri %98 ile word level vektörleştirme iken en zayıf performans ise %88 ile n gram vektörleştirme işlemi olmuştur.

3.3 Doğrusal Olmayan DVM İle Modellerin Test Edilmesi

Tüm modellerin doğruluk değerlerine baktığımızda en iyi soucu veren model Word Level olmaktadır. Ayrıca Word Level'ın farklı çekirdeklerde nasıl çalıştığını inceleyelim.

3.3.1 Hiperbolik Tanjant (Sigmoid) Çekirdeği

Hiperbolik Tanjant (Sigmoid) Çekirdeği, Sigmoid Çekirdeği ve Çok Katmanlı Algılayıcı (MLP) çekirdeği olarak da bilinir. Sigmoid çekirdeği aşağıdaki formül ile ifade edilir:

$$k(x, y) = tanh(\gamma.x^T y + r)$$

Sigmoid; yapay nöronlar için bir aktivasyon işlevi olarak kullanılan sinir ağının iki katmanlı, algılayıcı modeline eşdeğerdir.Bu çekirdek fonksiyonunun pratikte iyi performans gösterdiği tespit edilmiştir [11].

Bag Of Words: Bag Of Words'ü sigmoid çekirdeği ile test ettik ve sınıflandırma raporu;

Tablo 5: Bag Of Words Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.94	0.95	0.94	170
Magazin	0.99	0.99	0.99	152
Sağlık	0.99	0.97	0.98	151
Siyaset	0.96	0.97	0.96	146
Spor	1.00	0.94	0.97	148
Teknoloji	0.92	0.99	0.95	132
Yaşam	0.99	0.99	0.99	151
accuracy			0.97	1050
macro avg	0.97	0.97	0.97	1050
weighted avg	0.97	0.97	0.97	1050

şeklindedir. Çekirdeği değiştirdiğimizde modelin doğruluk oranı %97 olarak görülmektedir.

Word Level: Word Level'ı sigmoid çekirdeği ile test ettik ve sınıflandırma raporu;

Tablo 6: Word Level Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.97	0.96	0.96	170
Magazin	0.97	1.00	0.98	152
Sağlık	0.99	0.97	0.98	151
Siyaset	0.98	0.97	0.98	146
Spor	0.99	0.97	0.98	148
Teknoloji	0.96	0.98	0.97	132
Yaşam	0.99	0.99	0.99	151
accuracy			0.98	1050
macro avg	0.98	0.98	0.98	1050
weighted avg	0.98	0.98	0.98	1050

şeklindedir. Çekirdeği değiştirdiğimizde modelin doğruluk oranı %98 olarak görülmektedir.

N Gram: N Gram'ı sigmoid çekirdeği ile test ettik ve sınıflandırma raporu;

Tablo 7: N Gram Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.96	0.74	0.84	170
Magazin	0.92	0.90	0.91	152
Sağlık	0.60	0.96	0.74	151
Siyaset	0.96	0.92	0.94	146
Spor	0.97	0.89	0.93	148
Teknoloji	0.84	0.85	0.84	132
Yaşam	0.99	0.78	0.87	151
accuracy			0.86	1050
macro avg	0.89	0.86	0.87	1050
weighted avg	0.89	0.86	0.87	1050

şeklindedir. Çekirdeği değiştirdiğimizde modelin doğruluk oranı %86 olarak görülmektedir.

Char Level: Char Level'ı sigmoid çekirdeği ile test ettik ve sınıflandırma raporu;

Tablo 8: Char Level Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.95	0.94	0.95	170
Magazin	0.96	0.98	0.97	152
Sağlık	0.97	0.95	0.96	151
Siyaset	0.96	0.97	0.97	146
Spor	0.99	0.96	0.97	148
Teknoloji	0.95	0.95	0.95	132
Yaşam	0.96	0.98	0.97	151
accuracy			0.96	1050
macro avg	0.96	0.96	0.96	1050
weighted avg	0.96	0.96	0.96	1050

şeklindedir. Çekirdeği değiştirdiğimizde modelin doğruluk oranı %96 olarak görülmektedir.

3.3.2 Gaussian Radyal Tabanlı Fonksiyon Çekirdeği (RBF)

Ayarlanabilir parametre sigma, çekirdeğin performansında önemli bir rol oynar ve eldeki soruna göre dikkatle seçilmelidir. Aşırı tahmin söz konusu olduğunda, üstel ifade yaklaşık olarak doğrusallaşır ve yüksek boyutlu projeksiyon doğrusal olmayan yapısını kaybetmeye başlar. Öte yandan, eğer gerçek değerden daha düşük tahmin gerçekleşiyorsa, fonksiyon düzgünleştirmeyecek ve karar sınırı ise eğitim verisindeki gürültü değerlere karşı son derece

duyarlı olacaktır. RBF çekirdeği şu formül ile ifade edilir [11]:

$$k(x, y) = exp(-\frac{||x - y||^2}{2\sigma})$$

Bag Of Words: Bag Of Words'ü rbf çekirdeği ile test ettik ve sınıflandırma raporu;

Tablo 9: Bag Of Words Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.96	0.94	0.95	170
Magazin	0.97	0.99	0.98	152
Sağlık	0.97	0.96	0.96	151
Siyaset	0.97	0.97	0.97	146
Spor	0.99	0.95	0.97	148
Teknoloji	0.92	1.00	0.96	132
Yaşam	0.99	0.97	0.98	151
accuracy			0.97	1050
macro avg	0.97	0.97	0.97	1050
weighted avg	0.97	0.97	0.97	1050

şeklindedir. Çekirdeği değiştirdiğimizde modelin doğruluk oranı %97 olarak görülmektedir.
Word Level: Word Level'ı rbf çekirdeği ile test ettik ve sınıflandırma raporu;

Tablo 10: Word Level Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.99	0.95	0.97	170
Magazin	0.97	1.00	0.98	152
Sağlık	0.99	0.98	0.98	151
Siyaset	0.97	0.99	0.98	146
Spor	0.99	0.98	0.99	148
Teknoloji	0.98	1.00	0.99	132
Yaşam	0.99	0.98	0.98	151
accuracy			0.98	1050
macro avg	0.98	0.98	0.98	1050
weighted avg	0.98	0.98	0.98	1050

şeklindedir. Çekirdeği değiştirdiğimizde modelin doğruluk oranı %98 olarak görülmektedir.
N Gram: N Gram'ı rbf çekirdeği ile test ettik ve sınıflandırma raporu;

Tablo 11: N Gram Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	1.00	0.63	0.77	170
Magazin	0.93	0.90	0.92	152
Sağlık	0.87	0.85	0.86	151
Siyaset	0.97	0.90	0.94	146
Spor	0.98	0.86	0.91	148
Teknoloji	0.47	0.96	0.63	132
Yaşam	0.98	0.72	0.83	151
accuracy			0.83	1050
macro avg	0.89	0.83	0.84	1050
weighted avg	0.89	0.83	0.84	1050

şeklindedir. Çekirdeği değiştirdiğimizde modelin doğruluk oranı %83 olarak görülmektedir.

Char Level: Char Level'ı rbf çekirdeği ile test ettik ve sınıflandırma raporu;

Tablo 12: Char Level Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.97	0.95	0.96	170
Magazin	0.94	1.00	0.97	152
Sağlık	0.98	0.95	0.96	151
Siyaset	0.96	0.97	0.96	146
Spor	0.99	0.97	0.98	148
Teknoloji	0.96	0.97	0.96	132
Yaşam	0.97	0.97	0.97	151
accuracy			0.97	1050
macro avg	0.97	0.97	0.97	1050
weighted avg	0.97	0.97	0.97	1050

şeklindedir. Çekirdeği değiştirdiğimizde modelin doğruluk oranı %97 olarak görülmektedir.

4 YAPAY SİNİR AĞLARI İLE MODEL BELİRLEME

4.1 Yapay Sinir Ağları

Yapay sinir ağları biyolojik sinirlerin matematiksel modelinin genelleştirilmesi olarak da tanımlanabilir.

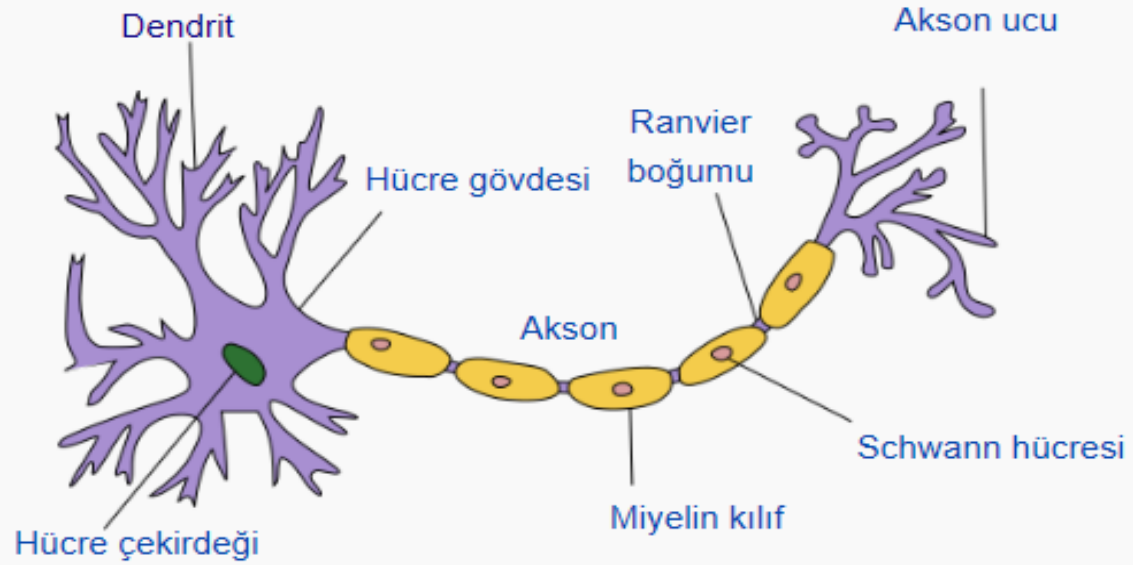
İlk yapay sinir agğ odeli 1943 yılında Warren McCulloch ve Walter Pitts tarafindan gerçekleştirilmişir. Yapay sinir hücreleri ile her türlü mantısal ifadeyi formülize etmenin mümkün olduğunu göstermişlerdir. Hücrelerin birbiri ile paralel çalışması gerektiği fikrini ortaya atarak öğrenme kurallarını belirlemeye başlamışlardır.

Yapay sinir ağları, insan beyninin özelliklerinden olan öğrenme yolu ile yeni bilgiler türetebilen bilgisayar sistemleridir. Bu yetenekleri geleneksel programlama yöntemleri ile gerçeklestirmek oldukça zordur. O nedenle, yapay sinir ağlarının, programlanması çok zor ya da mümkün olmayan olaylar için geliştirilmiş bilgi işleme ile ilgilenen bir bilgisayar bilim dalı olduğu söylenebilir.[12]

4.2 Yapay Sinir Ağlarının Yapısı

4.2.1 Biyolojik Sinir Hücresinin Özellikleri

Canlıların davranışlarını inceleyip, matematiksel olarak modelleyip, benzer yapay modellerin üretilmesine sibernetik denir. Eğitilebilir, adaptif ve kendi kendine organize olup öğrenebilen ve değerlendirme yapabilen yapay sinir ağları ile insan beyninin öğrenme yapısı modellenmeye çalışılmaktadır. Aynı insanda olduğu gibi yapay sinir ağları vasıtasıyla makinelerin eğitilmesi, öğrenmesi ve karar vermesi amaçlanmaktadır.
İnsandaki bir sinir hücresinin (nöron) yapısı şu şekildedir:



Şekil 20: Bir Sinir Hücresinin Biyolojik Gösterimi

Akson: Çıkış darbelerinin üretildiği elektriksel aktif gövdedir ve gövde üzerinde iletim tek yönlüdür. Sistem çıkışıdır.

Dentritler: Diğer hücrelerden gelen işaretleri toplayan elektriksel anlamda pasif kollardır. Sistem girişidir.

Sinaps: Hücrelerin aksonlarının diğer dentritlerle olan bağlantısını sağlar.

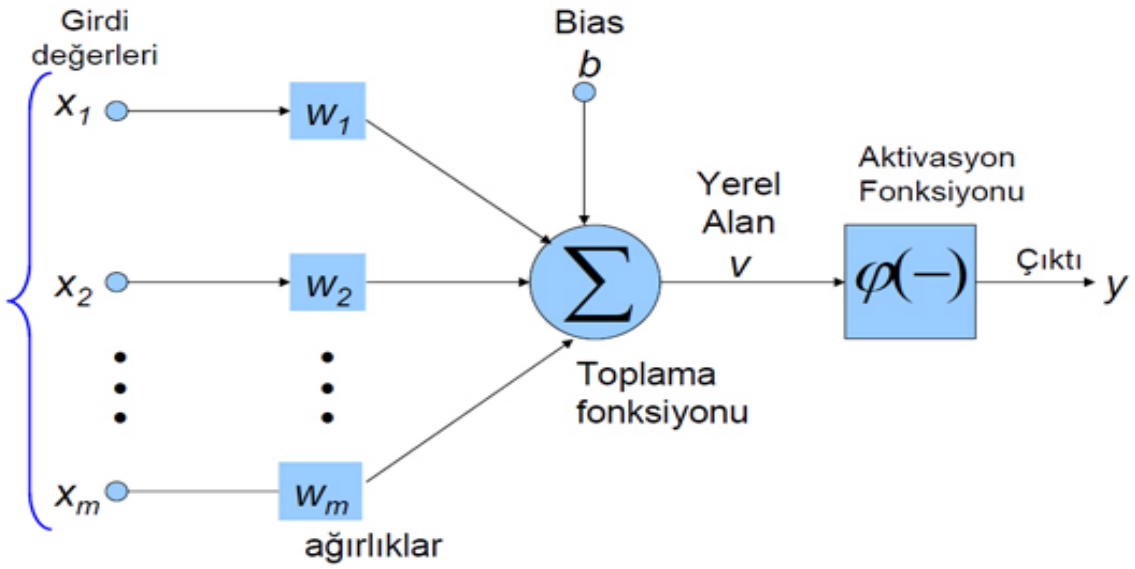
Miyelin Kılıf: Yayılma hızına etki eden yalıtım malzemesidir.

Çekirdek: Akson boyunca işaretlerin periyodik olarak yeniden üretilmesini sağlar.

Aksonda taşınan işaret sinapslara kimyasal taşıyıcılar yardımıyla iletilmektedir.Belirli bir eşik gerilim değerinin üstünde iken hücre uyarılırken, diğer durumlarda hücre bastırılır. Bu duruma göre çıkış işareti üretilmesine sinirsel hesaplama denir. [13]

4.2.2 Yapay Sinir Hücresinin Özellikleri

Biyolojik sinir ağlarının sinir hücreleri olduğu gibi yapay sinir ağlarının da yapay sinir hücreleri vardır. Yapay sinir hücreleri mühendislik biliminde işlem elemanları olarak da adlandırılmaktadır.



Şekil 21: Bir Sinir Hücresinin Matematiksel Modeli

Perseptron (Perceptron): Yapay sinir ağının en küçük parçası olarak bilinen perceptron, aşağıdaki gibi lineer bir fonksiyonla ifade edilmektedir ve ilk defa 1957 yılında Frank Rosenblatt tarafından tanımlanmıştır.
x_i bağımsız değişken, girdi olarak tanımlanır. Bu girdilerin her biri ağırlık w_i ile çarpılır. Ardından bu çarpım b ile toplanır ve

$$\sum x_i w_i + b$$

toplama fonksiyonu elde edilir. Bulunan bu toplama fonksiyonundan gelen değerler aktivasyon fonksiyonu ile belirli bir aralığa normalize edilir. Ardından da y_i çıkışı alınır. [13]

Yapay sinir hücresinin yapısını inceleyecek olursak;

Girdi: Girdiler tarafından bir yapay sinir hücresine bir başka yapay sinir hücresinden veya dış dünyadan bilgi alışı yapılır. Bunlar ağın öğrenmesi istenen örnekler tarafından belirlenir.

Ağırlıklar: Ağırlıklar bir yapay sinir hücresinin girişleri tarafından alınan bilgilerin önemini ve hücre üzerinde etkisi gösteren uygun katsayılardır. Her bir giriş için bir ağırlık vardır. Bu ağırlığın büyük olması bu girişin önemli olduğu ya da ağırlığın küçük olması girişin önem-

siz olduğunu göstermez. Bir ağırlığın değerinin sıfır olması o ağ için en önemli olay olabilir. Eksi değerler de yine girişin önemsiz olduğunu göstermez. Ağırlığın artı ve eksi olması girişin etkisinin pozitif ya da negatif olduğunu gösterir. Ağırlıklar değişken ya da sabit olabilirler.

Toplama Fonksiyonu: Toplama işlevi bir yapay sinirdeki her bir girdi ile o girdiye ait olan ağırlığın çarpılması, bu çarpımlara biasın eklenmesi ile hepsinin toplanmasıdır. Ayrıca her model ve her uygulama için bu toplama fonksiyonun kullanılması şart değildir. Bazı modeller, kullanılacak toplama fonksiyonunu kendileri belirler. Çoğu zaman daha karmaşık olan değişik toplama fonksiyonları kullanılır.

Aktivasyon Fonksiyonu: Yapay nöronun davranışını belirleyen önemli bir etken aktivasyon fonksiyonudur. Buna aynı zamanda "öğrenme eğrileri" de denir. Aktivasyon fonksiyonu hücreye gelen net girdiyi, diğer bir deyisle toplama fonksiyonunu işleyerek bu hücreye gelen girişlere karşılık olan çıkışı belirler.

Aktivasyon fonksiyonu da yapay sinir ağlarının farklı modelleri için farklı olabilir. En uygun aktivasyon fonksiyonunu belirlemek için geliştirilmiş bir fonksiyon yoktur. Toplama fonksiyonuna benzer sekilde hücrelerin hepsi için aynı aktivasyon fonksiyonu kullanma zorunlulugu yoktur. Bazıları aynı aktivasyon fonksiyonunu kullanırken bazıları kullanmayabilir. En çok kullanılan aktivasyon fonksiyonları sunlardır:

- Doğrusal Fonksiyon
- Basamak Fonksiyonu
- ReLU (Rectified Linear Unit-Düzeltilmiş Doğrusal Birim) Fonksiyonu
- Parçalı Dogrusal Fonksiyon
- Sigmoid Tipi Fonksiyon
- Tanjant Hiperbolik Tipli Fonksiyon
- Sinüs Tipli Fonksiyon

Aktivasyon Fonksiyonlarında Olması Gereken Özellikler İyi bir aktivasyon fonksiyonun sahip olması gereken bazı özellikler vardır.

1 -) Doğrusal Olmama: Aktivasyon fonksiyonları daha karmaşık problemlere çözüm bulabilmek için doğrusal olmamalıdır.

2 -) Türevlenebilirlik: Aktivasyon fonksiyonları sürekli olmalı ve bununla birlikte birinci dereceden türevi alınabilmelidir.

3 -) Aralık: Daha efektif bir model eğitimi için aktivasyon fonksiyonlarının değerinin sonsuza gitmesini değil, belirli aralıklar içinde olmasını isteriz.

4 -) Monotonluk: Aktivasyon fonksiyonlarının monoton olması minimum veya maksimum noktalarının olacağı anlamına gelir ve bu da öğrenmenin olacağının göstergesidir.

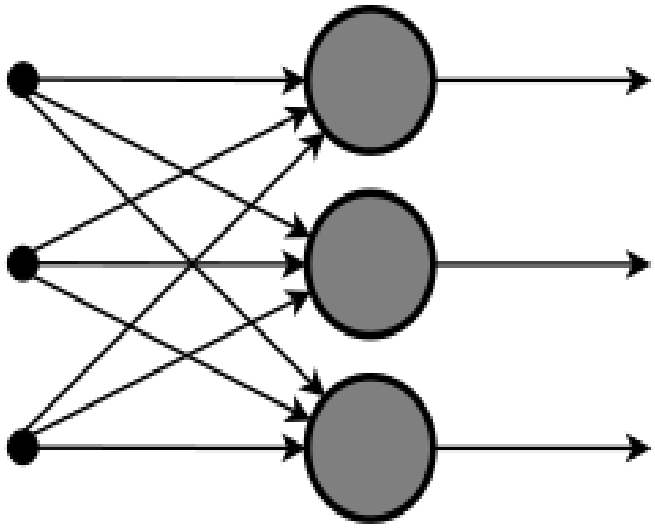
5 -) Orijine Göre Yakınsak: Aktivasyon fonksiyonları orijine göre yakınsak olduğunda YSA modellerinin başlangıç ağırlıkları rastgele küçük değerler olarak atandığında öğrenme gerçekleşir. Aksi takdirde daha farklı yöntemler ile modelin ağırlıkları atanmalıdır.[14]

Çıkış: Çıkış $y = f(v)$, aktivasyon fonksiyonunun sonucunun dış dünyaya veya diğer sinirlere gönderilmesidir. Bu sinirin çıkışı kendine ve kendinden sonra gelen bir ya da daha fazla sayıda sinire giriş olabilir. [12]

4.2.3 Yapay Sinir Ağı Yapısı

Yapay Sinir Ağları (YSA) tek katmanlı ve çok katmanlı olmak üzere iki sinir ağı modeline sahiptir.

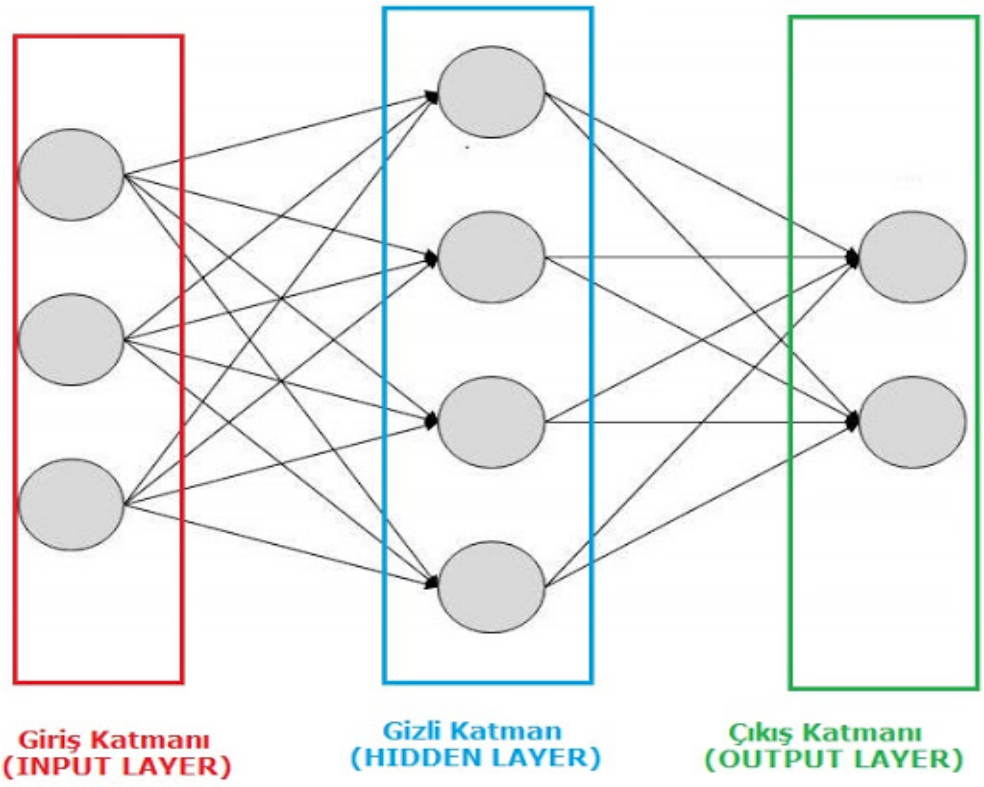
Tek Katmanlı YSA: Tek katmanlı algılayıcılardan oluşan yapay sinir ağları sadece girdi ve çıktı katmanlarından oluşur. Girdi katmanında giriş sinyalleri yer alırken çıkış katmanında ise algılayıcı nöronlar yer almaktadır. Giriş sinyalleri bütün çıkış nöronlarına bağlanmaktadır. Her bağlantının her çıkış nöronu için ayrı bir ağırlık değeri vardır.



Şekil 22: Tek Katmanlı Yapay Sinir Ağları

Doğrusal sınıflandırıcı (linear classifier) olarak kullanılan tek katmanlı algılayıcı ağlarda, çıkış fonksiyonu ikili (binary) değer üreten fonksiyondur. Çıkış değerleri kullanılan nöron modeline bağlı olarak {1,0} veya {-1,1} değerlerini almaktadır. Çıkış değerleri sınıfları temsil etmektedir. Ağa verilen giriş değerleri iki farklı sınıf arasında paylaştırılarak iki sınıfı birbirinden ayıran doğru veya düzlem bulunmaya çalışılır. Bu durum bir yapay sinir hücresinin sadece iki farklı sınıfı algılayabilmesini sağlamaktadır. İkiden fazla sınıfın algılanabilmesi için, birden çok yapay sinir hücresi kullanılması gerekmektedir. [15]

Çok Katmanlı YSA: Birden fazla yapay sinir hücresi bir araya gelerek yapay sinir ağını oluşturur. Bu hücrelerin bir araya gelmesi rastgele olmaz. Ağdaki sinir hücreleri katmanlar içine yerleştirilir. Yapay sinir ağları üç ana katmanda incelenir; Giriş Katmanı, Ara (Gizli) Katmanlar ve Çıkış Katmanı.



Şekil 23: Yapay Sinir Ağları Katmanları

Girdi Katmanı: Yapay sinir ağında gelen bilgiler girdi katmanında temsil edilir. Veri setine göre değişen özelliklerin her biri farklı bir düğüm olarak girdi katmanında temsil edilir. Her bir girdinin (özelliğin) ağırlık(weight) değeri vardır. Bu girdiler gizli katmandaki düğümler ile bu ağırlıklar aracılığı ile bağlıdır.

Gizli Katmanlar: Gizli katmanlar, girdi katmanından gelen bilgilerin işlenip bir çıktıya dönüştürüldüğü katmanlardır. Çıktıya dönüştürme işlemi yapay sinir ağının ağırlık değerleri kullanılarak gerçekleştirilir. Gizli katman sayısı problemin zorluğuna göre çeşitlilik gösterebilmektedir.

Çıktı Katmanı: Gizli katmandan gelen sonuç bilgisi hesaplanan çıktı değeri olarak belirlenir. Çıktı değeri veya değerlerinin tutulduğu katmandır. Verilen girdi değerine ait çıktıyı barındırır. [16]

4.3 YSA İle Modellerin Test Edilmesi

Yapay sinir ağları ile sınıflandırma yapmak için ilk olarak bir standartlaştırma işlemi yaparız. Bu standartlaştırmayı çoğu algoritma sever fakat YSA için standartlaştırma işlemi çok daha önemlidir. Bunun sebebi ise şöyle açıklanabilir YSA içinde birçok katman ve hücre bulunmaktadır, buralarda bulunan aykırılıklar ve istatistiksel varyans yapılarının birbirinden çok farklı olması elde edilecek sonuçların güvenilirliğini azaltır. Bu yüzden de YSA'da değişken standartlaştırma işlemi özellikle yapılır. Farklı standartlaştırma çeşitleri bulunmaktadır. Bu veri seti için **Standardizasyon** işlemini kullandık [17]. Standardizasyon, ortalama değerin 0, standart sapmanın ise 1 değerini aldığı, dağılımın normale yaklaştığı bir metoddur. Formülü şu şekildedir, elimizdeki değerden ortalama değeri çıkartıyoruz, sonrasında varyans değerine bölüyoruz [18]. Ardından **skelarn.neuralnetwork** modülünden çok katmanlı algılayıcı sınıflandırıcı MLPClassifier'ı kullanacağız.

4.3.1 Default Tanımlı YSA İle Modellerin Test Edilmesi

MLP'nin kendi içerisinde default olarak tanımlı olan parametreler şöyledir; aktivasyon fonksiyonu relu fonksiyonudur, girdi,gizli ve çıktı olmak üzere 3 katmanlıdır, gizli katmanın kendi içerisindeki boyut sayısı (hücre sayısı) 100'dür, ayrıca alpha değeri de 0.0001'dir.

Kısaca **ReLU** aktivasyon fonksiyonundan bahsedersek; fonksiyonunun ana avantajı aynı anda tüm nöronları aktive etmemesidir. Yani bir nöron negatif değer üretirse, aktive edilmeyeceği anlamına gelir.

$$g(x) = max(0, x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

şeklinde formülize edelir. Görebileceğimiz üzere negatif değerler üreten nöronlar sıfır değerini alır.Bu durum, ReLU'nun Hiperbolik Tanjant ve Sigmoid fonksiyonundan daha verimli ve hızlı çalışmasını sağlar. Bu nedenle ReLU, çok katmanlı sinir ağlarında daha çok tercih edilir.

Tüm vektörleştirme çeşitlerine bu YSA modelini uyguladık ve aşağıdaki sonuçları elde ettik:

Bag Of Words: Bag Of Words vektörleştirmesini YSA ile modellersek sınıflandırma raporu;

Tablo 13: Bag Of Words Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.93	0.94	0.94	170
Magazin	0.97	0.97	0.97	152
Sağlık	0.90	0.95	0.93	151
Siyaset	0.96	0.94	0.95	146
Spor	0.97	0.95	0.96	148
Teknoloji	0.98	0.98	0.98	132
Yaşam	0.97	0.93	0.95	151
accuracy			0.95	1050
macro avg	0.95	0.95	0.95	1050
weighted avg	0.95	0.95	0.95	1050

şeklindedir. Burada görüldüğü üzere modelin doğruluk değeri %95'dir.

Word Level: Word Level vektörleştirmesini YSA ile modellersek sınıflandırma raporu;

Tablo 14: Word Level Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.98	0.94	0.96	170
Magazin	0.93	0.99	0.96	152
Sağlık	0.96	0.96	0.96	151
Siyaset	0.96	0.94	0.95	146
Spor	0.97	0.97	0.97	148
Teknoloji	0.93	0.98	0.95	132
Yaşam	0.97	0.92	0.95	151
accuracy			0.96	1050
macro avg	0.96	0.96	0.96	1050
weighted avg	0.96	0.96	0.96	1050

şeklindedir. Burada görüldüğü üzere modelin doğruluk değeri %96'dır.

N Gram: N Gram vektörleştirmesini YSA ile modellersek sınıflandırma raporu;

Tablo 15: N Gram Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.92	0.78	0.84	170
Magazin	0.91	0.94	0.93	152
Sağlık	0.67	0.95	0.78	151
Siyaset	0.90	0.92	0.91	146
Spor	0.93	0.89	0.91	148
Teknoloji	0.93	0.83	0.88	132
Yaşam	0.97	0.80	0.88	151
accuracy			0.87	1050
macro avg	0.89	0.87	0.88	1050
weighted avg	0.89	0.87	0.87	1050

şeklindedir. Burada görüldüğü üzere modelin doğruluk değeri %87'dir.
Char Level: Char Level vektörleştirmesini YSA ile modellersek sınıflandırma raporu;

Tablo 16: Char Level Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.97	0.93	0.95	170
Magazin	0.96	0.97	0.97	152
Sağlık	0.92	0.95	0.93	151
Siyaset	0.96	0.97	0.96	146
Spor	0.99	0.97	0.98	148
Teknoloji	0.93	0.97	0.95	132
Yaşam	0.94	0.92	0.93	151
accuracy			0.95	1050
macro avg	0.95	0.95	0.95	1050
weighted avg	0.95	0.95	0.95	1050

şeklindedir. Burada görüldüğü üzere modelin doğruluk değeri %95'dir.

4.3.2 Yeni Tanımlı YSA İle Modellerin Test Edilmesi

Bu model için aktivasyon fonksiyonunu yine relu seçelim. Fakat bu sefer gizli katman sayısını arttıralım ve 3'e çıkaralım.Böylece toplam katman sayımız 5 olur. Gizli katmanların her birinin boyut sayısnı 100 yapalım. Her bir vektörleştirme işlemini bu model ile test ettik ve çıkan sınıflandırma raporlarına bakalım;
Bag Of Words: Bag Of Words vektörleştirmesini YSA ile modellersek sınıflandırma raporu;

Tablo 17: Bag Of Words Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.95	0.95	0.95	170
Magazin	0.99	0.95	0.97	152
Sağlık	0.96	0.98	0.97	151
Siyaset	0.97	0.97	0.97	146
Spor	0.98	0.97	0.98	148
Teknoloji	0.94	0.98	0.96	132
Yaşam	0.97	0.96	0.96	151
accuracy			0.97	1050
macro avg	0.97	0.97	0.97	1050
weighted avg	0.97	0.97	0.97	1050

şeklindedir. Burada görüldüğü üzere modelin doğruluk değeri %97'dir.

Word Level: Word Level vektörleştirmesini YSA ile modellersek sınıflandırma raporu;

Tablo 18: Word Level Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.97	0.92	0.95	170
Magazin	0.96	0.95	0.96	152
Sağlık	0.94	0.96	0.95	151
Siyaset	0.97	0.97	0.97	146
Spor	0.97	0.97	0.97	148
Teknoloji	0.92	0.97	0.94	132
Yaşam	0.97	0.95	0.96	151
accuracy			0.96	1050
macro avg	0.96	0.96	0.96	1050
weighted avg	0.96	0.96	0.96	1050

şeklindedir. Burada görüldüğü üzere modelin doğruluk değeri %96'dır.

N Gram: N Gram vektörleştirmesini YSA ile modellersek sınıflandırma raporu;

Tablo 19: N Gram Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.91	0.76	0.83	170
Magazin	0.94	0.86	0.90	152
Sağlık	0.91	0.84	0.88	151
Siyaset	0.88	0.89	0.88	146
Spor	0.64	0.93	0.75	148
Teknoloji	0.84	0.85	0.84	132
Yaşam	0.94	0.84	0.89	151
accuracy			0.85	1050
macro avg	0.87	0.85	0.85	1050
weighted avg	0.87	0.85	0.85	1050

şeklindedir. Burada görüldüğü üzere modelin doğruluk değeri %85'dir.

Char Level: Char Level vektörleştirmesini YSA ile modellersek sınıflandırma raporu:

Tablo 20: Char Level Sınıflandırma Raporu

	precision	recall	f1-score	support
Ekonomi	0.93	0.92	0.93	170
Magazin	0.95	0.96	0.95	152
Sağlık	0.88	0.95	0.91	151
Siyaset	0.93	0.95	0.94	146
Spor	0.99	0.98	0.99	148
Teknoloji	0.95	0.96	0.95	132
Yaşam	0.99	0.90	0.94	151
accuracy			0.94	1050
macro avg	0.95	0.95	0.95	1050
weighted avg	0.95	0.94	0.94	1050

şeklindedir. Burada görüldüğü üzere modelin doğruluk değeri %94'tür.

5 MODEL DEĞERLENDİRME KRİTERLERİ VE MODEL SEÇİMİ

5.1 Model Değerlendirme Kriterleri

Çalışmada kullanılmış olan makine öğrenme yöntemlerinin performansı değerlendirmek için, ölçme yöntemlerinden olan ve yaygın olarak kullanılan karışıklık matrisinin bileşenlerinden üretilmiş ölçüm değerleri kullanılmıştır. Bu ölçüm değerleri; Doğruluk Oranı (Accuracy), F1 Skoru (F1 Score), Kesinlik (Precision) ve Duyarlılık (Recall) hesaplamalarıdır. Bunların sonuçlarını sınıflandırma raporlarından elde ettik.
Performans ölçüm değerleri:
Doğruluk Oranı (Accuracy); tüm test seti içerisinde doğru sınıflandırılmış haberlerin sayısı oranıdır.

$$\text{Doğruluk Oranı} = \frac{TP + TN}{TP + TN + FP + FN}$$

Kesinlik (Precision); pozitif tahmin edilen haberlerin kaçının doğru olduğunun oranıdır. Yanlış pozitif tahminin maliyeti yüksek olduğu zaman kullanılabilecek bir ölçümdür. Bu çalışma özelinde Kesinlik, filtrelenmek istenen bir haberin sınıfına ait olmayan bir haberin filtrelenmesinin maliyeti olarak düşünülebilir.

$$\text{Başarım Oranı} = \frac{TP}{TP + FP}$$

Duyarlılık (Recall);gerçekte pozitif sınıfa ait olan haberin ne kadarının pozitif tahmin edildiği oranıdır. Yanlış negatif tahminin maliyeti yüksek olduğu zaman kullanılabilecek bir ölçümdür. Bu çalışmada özelinde Duyarlılık, filtrelenmek istenen bir haber sınıfına ait olan bir haberin filtrelenmemesinin maliyeti olarak düşünülebilir.

$$\text{Duyarlılık} = \frac{TP}{TP + FN}$$

F1 Skoru;kesinlik ve duyarlılığın harmonik ortalamasıdır. Her iki ölçümü aynı anda göz önünde bulundurmak açısından önemli bir göstergedir.

$$\text{F1 Skoru} = 2 * \frac{\text{Duyarlılık*Kesinlik}}{\text{Duyarlılık+Kesinlik}}$$

Tablo 21: Hata Matrisi Gösterimi

		Gerçekte Olan	
		Negatif (0)	Pozitif (1)
Tahmin Edilen	Negatif (0)	Doğru Negatif (True negative, TN): İkili sınıflandırmaprobleminde, 0 olarak tahmin edilen sınıfın gerçekte de 0 olması durumudur.	Yanlış Negatif (False negative, FN): İkili sınıflandırma probleminde, 0 olarak tahmin edilen sınıfın gerçekte 1 olması durumudur.
	Pozitif (1)	Yanlış Pozitif (False positive , FP): İkili sınıflandırma probleminde, 1 olarak tahmin edilen sınıfın gerçekte 0 olması durumudur.	Doğru Pozitif (True positive , TP): İkili sınıflandırma probleminde, 1 olarak tahmin edilen sınıfın gerçekte de 1 olması durumudur.

Diğer yandan bu çalışmada uygulanmış olan ikinci yöntem olan çok sınıflı sınıflandırmanın performansını ölçmek için doğruluk ölçüm değeri kullanılmıştır. Bu yaklaşımda doğruluk değeri; toplam doğru sınıflandırılmış haber sayısının, test setinde kullanılan toplam habere oranı ile elde edilmiştir. [19]

5.2 Test Veri Seti Hakkında

Test verisinde farklı haber sitelerinden alınmış her kategoriye ait birer haber başlığı seçtik. Test verimizin kategorisi sırayla 'Sağlık', 'Spor', 'Magazin', 'Siyaset', 'Ekonomi', 'Teknoloji', 'Yaşam' şeklindedir.

5.3 DVM Sınıflandırma Sonuçlarının Değerlendirilmesi

5.3.1 Lineer DVM

İlk olarak tüm metin vektörleştirme işlemlerini DVM'in lineer çekirdeği ile test edelim.

Tablo 22: Lineer DVM İçin Doğruluk Değerleri

	Metin Vektörleştirme Çeşitleri			
	Bag of Words	Word Level	N-Gram	Char Level
Lineer DVM Doğruluk Oranı	0.97	0.98	0.88	0.96

Linner DVM ve vektörleştirme çeşitlerine göre tüm test verilerini doğru sınıflandırma tahmini yapan vektörler Bag Of Words, Word Level ve Char Level'dır.

- Bag Of Words, Word Level, Char Level=['Sağlık', 'Spor', 'Magazin', 'Siyaset', 'Ekonomi', 'Teknoloji','Yaşam'] olacak şekilde hepsi doğru sınıflandırılmışltır.
- N Gram=['Sağlık', 'Spor', 'Teknoloji', 'Siyaset', 'Ekonomi', 'Teknoloji', 'Yaşam'] şeklinde sınıflandırma yapmıştır. Magazin haberini teknoloji olarak sınıflandırmıştır.

5.3.2 Sigmoid Çekirdeği İle DVM

DVM'i sigmoid çekirdeği ile test ettik.

Tablo 23: Sigmoid DVM İçin Doğruluk Değerleri

	Metin Vektörleştirme Çeşitleri			
	Bag of Words	Word Level	N-Gram	Char Level
Sigmoid DVM Doğruluk Oranı	0.97	0.98	0.86	0.96

Sigmoid çekirdeği ve vektörleştirme çeşitlerine göre tüm test verimizi doğru sınıflandırma tahmini yapabilen vektörleştirme işlemleri Bag of Words, Word Level ve Char Level olmaktadır.

- Bag Of Words, Word Level, Char Level=['Sağlık', 'Spor', 'Magazin', 'Siyaset', 'Ekonomi', 'Teknoloji','Yaşam'] olacak şekilde hepsi doğru sınıflandırılmışltır.

- N Gram=['Sağlık', 'Spor', 'Teknoloji', 'Siyaset', 'Ekonomi', 'Teknoloji', 'Yaşam'] şeklinde sınıflandırma yapmıştır. Magazin haberini teknoloji olarak sınıflandırmıştır. Lineer çekirdek için de N-Gram aynı şekilde çalışmıltır.

5.3.3 RBF Çekirdeği İle DVM

DVM'i rbf çekirdeği ile test ettik.

Tablo 24: Sigmoid DVM İçin Doğruluk Değerleri

	Metin Vektörleştirme Çeşitleri			
	Bag of Words	Word Level	N-Gram	Char Level
RBF DVM Doğruluk Oranı	0.97	0.98	0.83	0.97

Bu çekirdeği kullanarak test verimizi tahmin ettiğimizde Bag Of Words, Word Level ve Char Level test verimizdeki bütün habeleri doğru tahmin etmektedir.

- Bag Of Words, Word Level, Char Level=['Sağlık', 'Spor', 'Magazin', 'Siyaset', 'Ekonomi', 'Teknoloji','Yaşam'] olacak şekilde hepsi doğru sınıflandırılmışltır.

- N Gram=['Sağlık', 'Spor', 'Teknoloji', 'Siyaset', 'Ekonomi', 'Teknoloji', 'Yaşam'] şeklinde sınıflandırma yapmıştır. Magazin haberini teknoloji olarak sınıflandırmıştır. Lineer çekirdek için de N-Gram aynı şekilde çalışmıltır.

Görüldüğü üzere N-Gram hiçbir çekirdek için sınıflandırmanın tamamını doğru yapamamıştır. Diğer vektörleştirme çeşitlerinin doğruluk değerleri oldukça yüksektir ve test verimizin tüm haberlerini doğru sınıflandırmayı başarmışlardır.

5.4 YSA Sınıflandırma Sonuçlarının Değerlendirilmesi

5.4.1 Default YSA

Tablo 25: Default YSA İçin Doğruluk Değerleri

	Metin Vektörleştirme Çeşitleri			
	Bag of Words	Word Level	N-Gram	Char Level
Default YSA Doğruluk Oranı	0.96	0.95	0.88	0.95

Tabloda verilen vektörleştirme çeşitlerinden 1 yanlış sınıflandırma ile Bag Of Words, Default YSA için en iyi sınıflandırma yöntemi olmaktadır.

- Bag Of Words=['Sağlık', 'Spor', 'Teknoloji', 'Siyaset', 'Ekonomi', 'Teknoloji', 'Yaşam'] olacak şekilde sınıflandırmıştır.
- Word Level=['Sağlık', 'Spor', 'Siyaset', 'Siyaset', 'Ekonomi', 'Siyaset', 'Yaşam']
- N Gram=['Sağlık', 'Sağlık', 'Sağlık', 'Sağlık', 'Sağlık', 'Sağlık', 'Sağlık']
- Char Level=['Yaşam', 'Yaşam', 'Yaşam', 'Yaşam', 'Yaşam', 'Teknoloji', 'Yaşam']

Görüldüğü üzere en iyi sınıflandırmayı Bag Of Words yapmıştır.

5.4.2 Yeni Tanımlı YSA

Tablo 26: Default YSA İçin Doğruluk Değerleri

	Metin Vektörleştirme Çeşitleri			
	Bag of Words	Word Level	N-Gram	Char Level
Default YSA Doğruluk Oranı	0.97	0.96	0.85	0.94

Bu yeni tanımlı YSA ile vektörlerin sınıflandırmaları şu şekilde olmaktadır;

- Bag Of Words=['Sağlık', 'Spor', 'Teknoloji', 'Siyaset', 'Ekonomi', 'Teknoloji', 'Yaşam'] olacak şekilde sınıflandırmıştır.
- Word Level=['Teknoloji', 'Siyaset', 'Teknoloji', 'Siyaset', 'Ekonomi', 'Teknoloji', 'Yaşam']
- N Gram=['Sağlık', 'Spor', 'Spor', 'Siyaset', 'Siyaset', 'Spor', 'Spor']
- Char Level=['Ekonomi', 'Ekonomi', 'Ekonomi', 'Ekonomi', 'Ekonomi', 'Teknoloji', 'Yaşam']

Default YSA'da olduğu gibi 1 yanlış ile en iyi sınıflandırmayı Baf Of Words yapmıştır.

5.5 Model Seçimi

Genel olarak incelediğimizde DVM'in YSA'ya göre çoğu çekirdek ve vektörleştirme yöntemiyle daha iyi çalıştığı görülmektedir. DVM içinden de tüm vektörleştirme işlemlerinde en iyi sınıflandırmayı yapan Bag Of Words, Word Level ve Char Level olmaktadır. Bu proje kapsamında Bag Of Words'ü model olarak seçebiliriz.
DVM'in YSA'dan daha iyi çalışmasının bir çok sebebi sıralanabilir. Bunlardan biri öğrenme süreci olabilir. Öğrenme Süreci: YSA, ağırlıkların öğrenme sürecinde geri yayılım algoritmasını kullanır. Bu süreç, verilerin tümüyle eğitilirken ağırlıkların güncellenmesini gerektirir. SVM ise, eğitim sürecinde destek vektörlerini belirlemek için bir optimizasyon problemi çözer. Eğitim verilerinin çoğu dışında kalan destek vektörleri kullanılır, bu da DVM'in öğrenme sürecini daha verimli hale getirir. Ayrıca **Hiper Düzlem Ayırma Teoremi** 'nin DVM'de kullanılması DVM'i daha avantajlı ve kullanılabilir kılmaktadır.
Hiperdüzlem ayırma teoremi, çok boyutlu uzaylarda hiperdüzlemlerin verilen noktaları ayırabilme yeteneğine dair bir teoremdir. Temel olarak, hiperdüzlem ayırma teoremi, verilen bir nokta kümesinin, bu noktaların ait olduğu sınıflara göre hiperdüzlemlerle ayrıştırılabilmesi için gerekli ve yeterli koşulları tanımlar. DVM'de kullandığımız hiperdüzlem kavramı da bu teoreme dayanmaktadır. Veriler birçok düzlem ile ayrılabilir fakat en iyisi sınıflara eşit mesafede bulunan hiperdüzlemdir. Ayrıca biliyoruz ki DVM'in daha iyi çalışmasının bir başka sebebi de metin madenciliği alanında iyi bir performans sergilemesidir.

6 UYGULAMA

Farklı haber sitelerinden aldığımız 50 haber başlığını Bag Of Words kullanarak sınıflandıralım. Hangi sınıflandırma yöntemi daha yüksek doğru ile çalışır karşılaştıralım. Bu haberlere daha önce veri setimize uyguladığımız işlemleri uyguladık ve tahminimiz o şekilde yaptık. Seçtiğimiz bazı haberleri şöyle gösterebiliriz;

Tablo 27: Uygulama İçin Kullanılan Haberlerin Bir Kısmı

HABERLER	ETİKET
Aslan devler sahnesine çıkıyor! İşte Galatasaray'ın Şampiyonlar Ligi'ndeki muhtemel rakipleri	Spor
Tamer Karadağlı'dan Merve Dizdar'ın Cannes'daki sözlerine eleştiri	Magazin
Son Dakika: Devlete olan borçların yapılandırılması için başvuru ve ilk taksit ödeme süresi uzatıldı	Ekonomi
Defne Devlet Hastanesinde 2 hafta içinde ameliyatlara başlanacak	Sağlık
Instagram'ın kurucu ortakları yeni bir sosyal haber ağı uygulaması sunuyor	Teknoloji
Gelecek Partili Selçuk Özdağ: 10 milletvekili arkadaş, CHP'ye istifa dilekçelerimizi verdik	Siyaset
Mutlu insanlar mı sebze yer yoksa sebze yemek mi mutlu eder? Hangi sebze ve meyveleri tüketmek daha etkili?	Yaşam
Filenin Sultanları, Voleybol Milletler Ligi'nde oynadığı ilk maçta Güney Kore'yi 3-0'lık set sonucuyla mağlup etti.	Spor
Türkiye Cumhuriyet Merkez Bankasının (TCMB) toplam rezervi geçen hafta 3 milyar 537 milyon dolar azalarak 101 milyar 590 milyon dolara geriledi. Net rezervler ise -115,3 milyon dolara indi.	Ekonomi
Boşanma iddiasıyla gündeme gelen Kıvanç Tatlıtuğ, eşiyle düşman çatlattı	Magazin
Fenerbahçe'nin Antalyaspor karşısında mücadeleyi 1-0 önde götürdüğü sırada Ankara'dan Galatasaray'ın gol yediği haberinin gelmesiyle Kadıköy adeta yıkıldı. Sarı-lacivertli taraftarlar tribünde bir anda "Gol" diye bağırırken, şampiyonluk tezahüratları yapılmaya başlandı.	Spor
Güçlü bacakları olan kişilerde kalp krizi sonrası kalp yetmezliği riski daha düşük	Sağlık
Meta'nın Ticaret ve Finansal Teknolojiler Başkanı Stephane Kasriel, Twitter'da şirketin NFT ve dijital koleksiyon özelliklerini Instagram ve Facebook'ta kullanımdan kaldıracağını duyurdu.	Teknoloji
Milletvekilleri yemin ederek görevlerine başlayacak: Meclis'in yeni dönem yemin töreni ne zaman?	Siyaset

Bu haberler için DVM ve YSA'da en iyi çalışan Bag Of Words vektörleştirmesini kullanarak sınıflandırdık.

DVM: Destek vektör makinesinde Bag Of Words'ü ve çekirdek olarak RBF çekirdeğini kullanarak tahmin ettik ve 50 adet haber başlığı için 9 tanesini yanlış sınıflandırmıştır. Bunu yüzdelik olarak yorumlarsak %82'sini doğru sınıflandırmıştır.

YSA: Yapay sinir ağlarında Bag Of Words'ü ve default tanımlı YSA'yı kullanarak tahmin ettik ve 50 adet haber başlığı için 14 tanesini yanlış sınıflandırmıştır. Bunu yüzdelik olarak yorumlarsak %72'sini doğru sınıflandırmıştır.

Model seçiminde de belirttiğimiz gibi Destek Vektör Makinleri sınıflandırma problemlerinde Yapay Sinir Ağlarına göre daha iyi çalışmaktadır. Bu problem için seçebileceğimiz en iyi sınıflandırma problemi Destek Vektör Makinesidir.

Kaynaklar

[1] O. USLU, S. AKYOL 2021. "Türkçe Haber Metinlerinin Makine Öğrenmesi Yöntemleri Kullanılarak Sınıflandırılması", ESTUDAM Bilişim Dergisi Cilt 2, Sayı 1, 15-20, 2021.

[2] D. Kılınç, E. Borandağ, F. Yücalar, V. Tunali, M. Şimşek, ve A. Özçift. 2016. "KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi". Marmara Fen Bilim. Derg., 28(3), 89–94.

[3] C. Toraman, F. Can, S. Koçberber. 2011. "Developing a Text Categorization Template for Turkish News Portals". 2011 International Symposium on Inovations in Intelligent Systems and Applications, 379–383.

[4] P. Tüfekci, E. Uzun, ve B. Sevinç. 2012. "Türkçe Dilbilgisi Özelliklerini Kullanarak Web Tabanlı Haber Metinlerinin Sınıflandırılması". 2012 20th Signal Processing and Communications Applications Conference (SIU), 1–4.

[5] Bai X. (2011) "Predicting consumer sentiments from online text". Decision Support Systems, 50(4), 732-742

[6] S. Camilleri, M. R. Agius, J. Azzopardi. 2020. "Analysis of Online News Coverage on Earthquakes Through Text Mining". Front. Earth Sci., 8(May), 1–12.

[7] Kaggle. Turkish Headlines Dataset. 2021. (Erişim Tarihi:15.11.2022)
`https://www.kaggle.com/datasets/anil1055/turkish-headlines-dataset`

[8] M. Ç. AKSU, E. KARAMAN, 2020. "FastText ve Kelime Çantası Kelime Temsil Yöntemlerinin Turistik Mekanlar İçin Yapılan Türkçe İncelemeler Kullanılarak Karşılaştırılması". Avrupa Bilim ve Teknoloji Dergisi Sayı 20, S. 311-320, Aralık 2020.

[9] D.KILINÇ 2018. Metin İşleme 1 — Eski Tarz Yöntemler (Bag of Words ve TFxIDF) (Erişim Tarihi:08.12.2022)
`https://medium.com/deep-learning-turkiye/metin-i%CC%87%C5%9Fleme-1-eski-tarz-y%C3%B6ntemler-bag-of-words-ve-tfxidf-76d5a0cf1b29`

[10] Kaynak Kullanılan Kod :
`https://github.com/denopas/TextProcessing/blob/master/TextProcessingPart1.ipynb`

[11] E.GÜLDOĞAN 2017. "Çeşitli Çekirdek Fonksiyonları İle Oluşturulan Destek Vektör Makinesi Modellerinin Performansının İncelenmesi: Bir Klinik Uygulama". İnönü Üniversitesi ve Mersin Üniversitesi Biyoistatistik Ve Tıp Bilişimi Anabilim Dalı Ortak Doktora Tezi, 2017

[12] Öztemel E., "Yapay Sinir Ağları", Papatya Yayıncılık, İstanbul, s.6-8, (2003).

[13] ŞU KARA KUTUYU AÇALIM: Yapay Sinir Ağları
https://ayyucekizrak.medium.com/%C5%9Fu-kara-kutuyu-a%C3%A7alim-yapay-sinir-a%C4%9Flar%C4%B1-7b65c6a5264a

[14] Aktivasyon Fonksiyonları
https://devreyakan.com/aktivasyon-fonksiyonlari/#:~:text=Aktivasyon%20fonksiyonlar%C4%B1%20n%C3%B6ronlar%C4%B1n%20toplam%20fonksiyonunda,lineer(do%C4%9Frusal)%20olamayan%20fonksiyonlard%C4%B1r.

[15] Yapay Sinir Ağları ve Tek Katmanlı Ağlarda Öğrenme
https://www.linkedin.com/pulse/yapay-sinir-a%C4%9Flar%C4%B1-ve-tek-katmanl%C4%B1-a%C4%9Flarda-%C3%B6%C4%9Frenme-tanju-do%C4%9Fan/?originalSubdomain=tr

[16] S.ÇALIŞKAN, S.A.YAZICIOĞLU, U.DEMİRCİ, Z. KUŞ, "YAPAY SİNİR AĞLARI, KELİME VEKTÖRLERİ VE DERİN ÖĞRENME UYGULAMALARI", Fatih Sultan Mehmet Vakıf Üniversitesi Mühendislik ve Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı, 2018 İstanbul, s 5-6.

[17] (50 Saat) Python A-Z: Veri Bilimi ve Machine Learning-Yapay Sinir Ağları Model,Video 303
https://www.udemy.com/course/python-egitimi/learn/lecture/14599938#overview

[18] Veri Hazırlığının Vazgeçilmezi : Özellik Ölçeklendirme
www.veribilimiokulu.com/veri-hazirliginin-vazgecilmezi-ozellik-olceklendirme/

[19] İ.ŞAHİN 2019. "METİN MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ İLE İNTERNET SAYFALARININ SINIFLANDIRILMASI". Hacettepe Üniversitesi Lisansüstü Egitim – Öğretim ve Sınav Yönetmeliği Endüstri Mühendisliği Anabilim Dalı YÜKSEK LİSANS TEZİ.

[20] Github Kodları
https://github.com/Hilalemin/SVM-ile-metin-siniflandirma