**TÜRKÇE HABER METINI TEMIZLEME VE SINIFLANDIRMA ÇALIŞMASI**

**ZEHRA BİRGE**

Öğrenci, Bitlis Eren Üniversitesi, Mühendislik ve Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, zbirge8@gmail.com

|  |
| --- |
| **ÖZET**  Bu çalışma, Türkçe haber verilerinin analizi ve işlenmesi üzerine odaklanmaktadır. Veri setimiz, toplamda 70.000 haber içermekte olup, bu haberler çeşitli kategorilere ayrılmaktadır. Kategoriler arasında Ekonomi, Spor, Sağlık, Teknoloji ve Eğitim gibi alanlar yer almakta, bu da çalışmanın geniş bir yelpazede konuları kapsamasını sağlamaktadır.  Çalışmanın ilk aşaması, ham verilerin analiz edilebilir bir forma dönüştürülmesi için çeşitli ön işleme tekniklerinin uygulanmasıdır. Bu süreç, metin temizleme, durak kelimelerin (stopwords) çıkarılması ve lemmatizasyon gibi adımları içermektedir. Temizlenmiş metinler, ardından TF-IDF (Term Frequency-Inverse Document Frequency) yöntemiyle vektörleştirilerek makine öğrenimi algoritmaları için uygun hale getirilmiştir. Bu aşamada, metinlerin anlamını koruyarak sayısal verilere dönüştürülmesi amaçlanmaktadır.  Veri setinin özellikleri çıkarıldıktan sonra, çeşitli makine öğrenimi algoritmaları kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. Bu bağlamda, Naive Bayes, Destek Vektör Makineleri (SVM) ve Lojistik Regresyon modelleri test edilmiştir. Deneyler sonucunda, en yüksek doğruluk oranı Naive Bayes modeli ile %85 olarak elde edilmiştir. Bu sonuç, Türkçe metinlerin sınıflandırılmasında Naive Bayes algoritmasının etkinliğini ortaya koymaktadır.  Bu çalışma, Türkçe doğal dil işleme alanında önemli bir katkı sunmayı hedeflemektedir. Uygulanan yöntemler ve elde edilen sonuçlar, hem akademik literatüre hem de pratik uygulamalara ışık tutmaktadır. Araştırmacılar için kapsamlı bir rehber niteliği taşıyan bu çalışma, Türkçe metin analizi ve sınıflandırması konusundaki mevcut bilgi birikimini genişletmektedir. Ayrıca, benzer projeler için bir temel oluşturarak, Türkçe doğal dil işleme alanında daha fazla araştırma yapılmasına teşvik etmektedir.  Gelecek çalışmalar, farklı makine öğrenimi algoritmalarının yanı sıra derin öğrenme yöntemlerini de içerebilir. Ayrıca, daha fazla veri seti ile yapılan deneyler, model performansını artırmak için yeni yaklaşımlar geliştirilmesine olanak tanıyacaktır. Bu bağlamda, Türkçe doğal dil işleme alanında daha geniş çaplı ve derinlemesine analizlerin yapılması beklenmektedir.  **Anahtar Kelimeler:** Türkçe metin işleme, TF-IDF, sınıflandırma, doğal dil işleme, veri seti analizi.  *.*  **ABSTRACT**  This study focuses on the analysis and processing of Turkish news data. Our dataset contains a total of 70,000 news items, which are categorized into various categories. The categories include areas such as Economy, Sports, Health, Technology and Education, which allows the study to cover a wide range of topics.  The first stage of the study is to apply various pre-processing techniques to transform the raw data into an analyzable form. This process includes steps such as text cleaning, stopwords removal and lemmatization. The cleaned texts are then vectorized using the TF-IDF (Term Frequency-Inverse Document Frequency) method to make them suitable for machine learning algorithms. At this stage, the aim is to convert the texts into numerical data while preserving their meaning.  After extracting the features of the dataset, classification was performed using various machine learning algorithms. In this context, Naive Bayes, Support Vector Machines (SVM) and Logistic Regression models were tested. As a result of the experiments, the highest accuracy rate of 85% was obtained with the Naive Bayes model. This result demonstrates the effectiveness of the Naive Bayes algorithm in the classification of Turkish texts.  This study aims to make a significant contribution to the field of Turkish natural language processing. The methods applied and the results obtained shed light on both academic literature and practical applications. As a comprehensive guide for researchers, this study extends the existing body of knowledge on Turkish text analysis and classification. It also provides a foundation for similar projects and encourages further research in the field of Turkish natural language processing.  Future work could include different machine learning algorithms as well as deep learning methods. Moreover, experiments with more datasets will allow the development of new approaches to improve model performance. In this context, more extensive and in-depth analyses in the field of Turkish natural language processing are expected.  **Keywords:** Turkish text processing, TF-IDF, classification, natural language processing, dataset analysis. |

**1. GİRİŞ**

Türkçe doğal dil işleme (NLP) çalışmaları, son yıllarda bilgiye erişim, içerik yönetimi ve metin analizi gibi alanlarda giderek daha fazla önem kazanmıştır. Bu bağlamda, haber metinlerinin sınıflandırılması, yalnızca kullanıcılara doğru bilgiye hızlı bir şekilde ulaşma imkânı sağlamakla kalmayıp, aynı zamanda içerik yönetiminde de büyük kolaylıklar sunmaktadır. Günümüzde dijital ortamda yayımlanan haberlerin sayısının artması, bu içeriklerin etkin bir şekilde yönetilmesini zorunlu hale getirmiştir. Kullanıcıların ilgi alanlarına göre özelleştirilmiş içerik sunulması, haber platformlarının rekabet avantajını artırırken, aynı zamanda doğru bilgiye erişimi de hızlandırmaktadır.

Ancak, Türkçe'nin eklemeli bir dil olması, bağlam bağımlılığı gibi dilin yapısal özellikleri, doğal dil işleme projelerinde karşılaşılan zorlukları artırmaktadır. Türkçe’deki kelime kökleri, eklemelerle zenginleştirildiği için, dilin morfolojik yapısının karmaşıklığı, metinlerin analizi ve sınıflandırılması sırasında önemli engeller oluşturmaktadır. Ayrıca, bağlam bağımlılığı, kelimelerin anlamlarının cümle içerisindeki kullanımlarına göre değişiklik göstermesi, doğru analiz ve sınıflandırma yapabilmek için daha karmaşık algoritmalar geliştirilmesini gerektirmektedir.

Türkçe haber metinlerinin sınıflandırılması, doğal dil işleme uygulamaları arasında kritik bir yere sahiptir. Örneğin, dijital haber platformları ve sosyal medya üzerindeki içeriklerin düzenlenmesi ve filtrelenmesi, bu tür sınıflandırma projeleri sayesinde mümkün olmaktadır. Bu sayede kullanıcılar, kendilerine en uygun ve ilgi çekici içeriklere daha hızlı ulaşabilmekte, aynı zamanda bilgi kirliliğinden korunabilmektedir. Bununla birlikte, Türkçe için yapılan doğal dil işleme çalışmaları, genellikle İngilizce gibi dillerle karşılaştırıldığında daha az kaynak ve araç ile desteklenmektedir. Bu durum, dilin yapısal özelliklerinden kaynaklanan zorlukların yanı sıra geniş veri setlerinin eksikliğinden dekaynaklanmaktadır. Türkçe’ye özgü dil işleme araçlarının ve kaynaklarının azlığı, araştırmacıların ve geliştiricilerin bualanda daha fazla çaba sarf etmesini zorunlu kılmaktadır.

Bu çalışma, 70.000 haber içeren geniş bir veri setini kullanarak Türkçe haber metinlerini sınıflandırma üzerine odaklanmıştır. Veri seti, haber başlıklarını, içeriklerini ve kategorilerini içermekte olup, Ekonomi, Spor, Sağlık, Teknoloji ve Eğitim gibi alanlarda zengin bir içerik sunmaktadır. Çalışma, verilerin ön işleme sürecinden başlayarak TF-IDF yöntemiyle özellik çıkarımı ve farklı makine öğrenimi algoritmalarıyla sınıflandırma adımlarını kapsamaktadır. Ayrıca, elde edilen sonuçlar, kategoriler arasındaki dengesizliklerin modeller üzerindeki etkisini de değerlendirmektedir. Bu bağlamda, çalışma, Türkçe doğal dil işleme alanında önemli bir katkı sunmayı amaçlamakta ve benzer projelere ışık tutmayı hedeflemektedir.

**2. YÖNTEM**

Bu bölümde, Türkçe haber metinlerinin sınıflandırılması amacıyla gerçekleştirilen çalışmanın yöntemleri detaylı bir şekilde ele alınacaktır. Veri setinin temininden başlayarak metin temizleme, özellik çıkarımı, modelleme süreçleri ve model performansını değerlendirmek için kullanılan ölçütler kapsamlı bir şekilde açıklanacaktır. Her bir adım, uygulanan teknikler ve bunların nedenleri ile birlikte sunulacaktır.

**2.1. Veri Seti**

Bu çalışmada kullanılan veri seti, Kaggle platformundan temin edilmiştir ve toplamda **70.000** haber metnini içermektedir. Veri seti, çeşitli kategorilere ait haberler barındırmakta olup, bu kategoriler arasında **Ekonomi (%20)**, **Spor (%15)**, **Sağlık (%10)**, **Teknoloji (%8)** ve **Eğitim (%7)** gibi alanlar bulunmaktadır. Bu çeşitlilik, modelin farklı konularda eğitilmesine olanak tanıyarak, genel performansını artırmaktadır.

Veri setinin yapısı aşağıdaki gibi tanımlanmıştır:

|  |  |
| --- | --- |
| **Sütun Adı** | **Açıklama** |
| |  |  | | --- | --- | | **id** |  | |  |  | | Her bir haber için benzersiz kimlik numarası. |
| **main\_image** | Haberin ana görseline ait URL. |
| **published** | Haberin yayımlanma tarihi ve saati. |
| **site** | Haberin yayımlandığı haber kaynağı. |
| **text** | Haberin tam metni |
| **title** | Haberin başlığı. |
| **url** | Haberin yayımlandığı web adresi. |

Metinlerin ortalama uzunluğu **250** kelime civarındadır ve metinler, **30'dan fazla** güvenilir haber kaynağından derlenmiştir. Bu çeşitlilik, veri setini geniş kapsamlı doğal dil işleme (NLP) çalışmaları için uygun bir kaynak haline getirmektedir. Ayrıca, veri setinin güncel olayları kapsaması, modelin gerçek dünya uygulamalarında daha etkili olmasını sağlayacaktır.

**2.2. Metin Temizleme**

Metin temizleme, ham verilerden anlamlı bilgiler çıkarma sürecinde kritik bir adımdır. Bu aşama, modelin performansını artırmak ve gereksiz gürültüyü azaltmak için oldukça önemlidir. Metin temizleme işlemleri, aşağıdaki adımlarla gerçekleştirilmiştir:

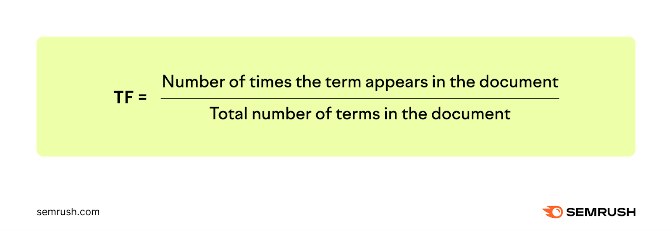
1. **Fazla Boşlukların Kaldırılması**: Metin içerisindeki gereksiz boşluklar ve satır sonları temizlenmiştir. Bu adım, veri setinin daha düzenli ve analiz edilebilir olmasını sağlamaktadır. Gereksiz boşluklar, kelimelerin ayrımını zorlaştırabilir ve modelin öğrenme sürecinde hatalara yol açabilir.
2. **Noktalama İşaretlerinin Kaldırılması**: Metin içerisindeki özel karakterler ve noktalama işaretleri çıkarılmıştır. Bu işlem, kelimelerin daha net bir şekilde tanımlanmasına yardımcı olmaktadır. Noktalama işaretlerinin kaldırılması, metinlerin daha homojen bir yapıya kavuşmasını sağlar.
3. **Durak Kelimelerin Çıkarılması**: Türkçe durak kelimeler (örneğin "bir", "ve", "de") metinden çıkarılmıştır. Durak kelimeler, anlam açısından zayıf olduklarından, bu adım modelin daha anlamlı kelime öbekleri ile çalışmasını sağlamaktadır. Durak kelimelerin çıkarılması, metinlerin anlamını korurken, modelin daha etkili bir şekilde öğrenmesini sağlar.
4. **Küçük Harfe Dönüştürme**: Tüm metin, tutarlılığı sağlamak amacıyla küçük harfe dönüştürülmüştür. Bu, kelimelerin eşit bir şekilde değerlendirilmesine olanak tanımaktadır. Büyük harf ve küçük harf ayrımının ortadan kaldırılması, kelimelerin daha tutarlı bir şekilde analiz edilmesini sağlar.
5. **Lemmatizasyon**: Kelimeler, köklerine indirgenerek anlam bütünlüğü korunmuştur. Bu işlem, kelimelerin çeşitli çekimlerine rağmen aynı anlamda değerlendirilmesini sağlamaktadır. Lemmatizasyon, dilin morfolojik yapısını göz önünde bulundurarak, kelimelerin köklerini bulmak için dilbilgisel kuralları kullanır. Bu sayede, farklı şekillerdeki kelimeler tek bir kök altında toplanır.

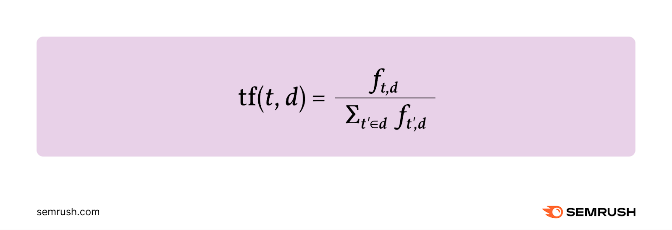
**2.3. TF-IDF Vektörleştirme**

TF-IDF (Term Frequency-Inverse Document Frequency) yöntemi, metinlerdeki kelimelerin önem derecesini hesaplayan bir tekniktir. TF-IDF'in temel amacı, sık kullanılan ancak düşük anlam taşıyan kelimelerin etkisini azaltarak daha anlamlı özellikler elde etmektir. TF-IDF, kelimelerin sıklığını ve belge içindeki dağılımını dikkate alarak her kelimeye bir ağırlık atar.

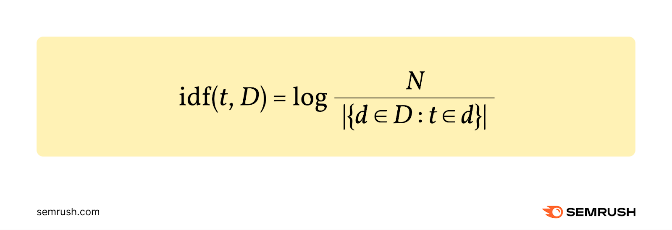
**TF-IDF Hesaplama Formülü:**

* **Term Frequency (TF)**: Bir terimin (kelimenin) belgede ne sıklıkta geçtiğini gösterir.



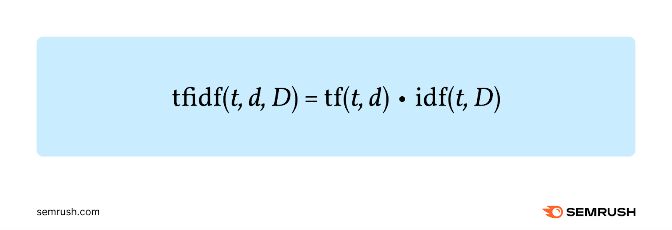
​

* **Inverse Document Frequency (IDF)**: Bir terimin ne kadar bilgilendirici olduğunu gösterir. Çok sayıda belgede bulunan kelimeler düşük IDF değerine sahipken, nadir bulunan kelimeler yüksek IDF değerine sahiptir.



Burada, N toplam belge sayısını, ∣{d∈D:t∈d}∣ ise terimin bulunduğu belge sayısını ifade eder.

* **TF-IDF**: Her kelimenin TF ve IDF değerlerinin çarpımıdır.

Bu çalışmada, TF-IDF yöntemiyle maksimum **5.000** özellik çıkarılmıştır. Özellik çıkarımı, modelin öğrenme sürecinde kritik bir rol oynamakta ve metinlerin daha iyi temsil edilmesini sağlamaktadır. TF-IDF, metinlerin anlamını korurken, önemli kelimeleri vurgular ve modelin odaklanmasını istediğimiz alanları belirler.

**2.4. Veri Ayırma ve Modelleme**

Veri setinin etkin bir şekilde kullanılması için, eğitim ve test verileri olarak ikiye ayrılması gerekmektedir. Bu ayrım, modelin genel performansını değerlendirmek için önemlidir. Veri seti, **%80 eğitim** ve **%20 test** olarak bölünmüştür. Bu oran, modelin öğrenme sürecinde yeterli veriye sahip olmasını sağlarken, test aşamasında da güvenilir sonuçlar elde edilmesine olanak tanır.

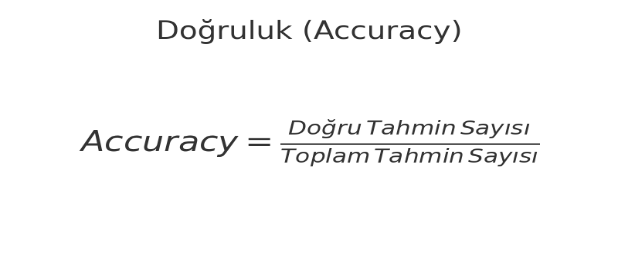
Kullanılan sınıflandırma modelleri şunlardır:

1. **Naive Bayes**: Probabilistik bir sınıflandırma algoritmasıdır. Özellikle metin sınıflandırma görevlerinde yaygın olarak kullanılmaktadır. Naive Bayes, bağımsızlık varsayımına dayanarak, her bir sınıf için olasılıkları hesaplar ve en yüksek olasılığa sahip sınıfı tahmin eder. Bu algoritma, hızlı ve etkili bir şekilde çalışabilmesi ile bilinir.
2. **Destek Vektör Makineleri (SVM)**: Çizgisel ve doğrusal olmayan sınıflandırma problemleri için uygun bir algoritmadır. SVM, verileri en iyi ayıran hiper düzlemi bulmaya çalışır ve bu sayede yüksek boyutlu veri setlerinde etkili bir şekilde çalışabilmektedir. SVM, genellikle yüksek doğruluk oranları ile sonuç verir.
3. **Lojistik Regresyon**: Özellikle doğrusal sınıflandırma problemlerinde yaygın olarak kullanılan bir yöntemdir. Lojistik regresyon, her bir sınıf için bir olasılık tahmini yaparak, en yüksek olasılığa sahip sınıfı belirler. Bu yöntem, basitliği ve yorumlanabilirliği ile tercih edilmektedir.

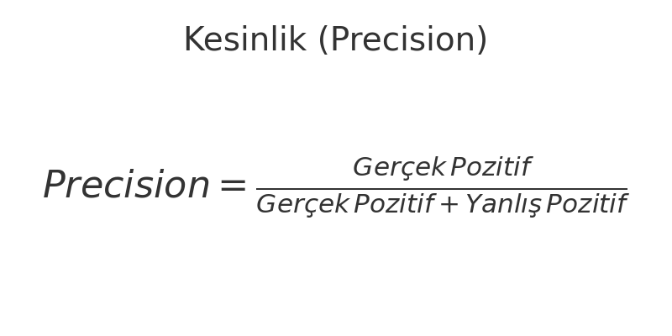
**2.5. Model Performansını Değerlendirme**

Modelin başarısını değerlendirmek için çeşitli ölçütler kullanılmaktadır. Bu ölçütler, modelin ne kadar doğru tahminler yaptığını ve hangi sınıflarda daha başarılı olduğunu anlamamıza yardımcı olur. Kullanılan ölçütler şunlardır:

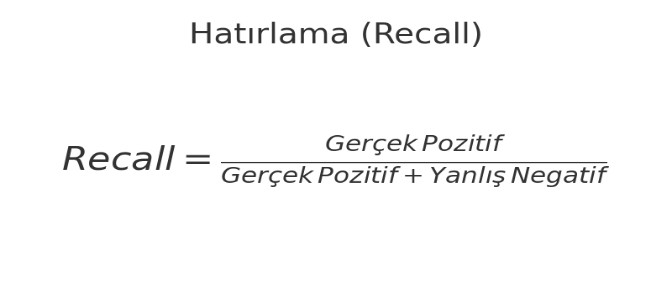
1. **Doğruluk (Accuracy)**: Modelin doğru tahmin ettiği örneklerin toplam örneklere oranıdır.

Doğruluk, genel bir performans ölçütü olsa da, dengesiz veri setlerinde yanıltıcı olabilir.

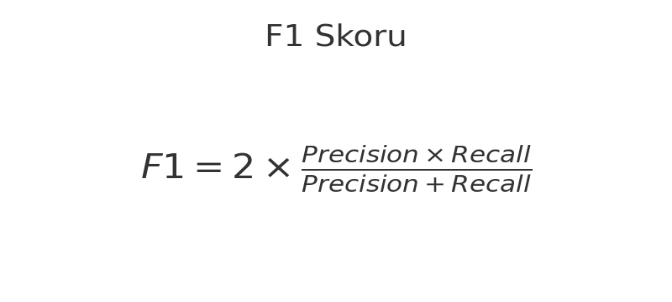
1. **Kesinlik (Precision)**: Modelin pozitif olarak tahmin ettiği örneklerden kaç tanesinin gerçekten pozitif olduğunu gösterir.

Kesinlik, özellikle yanlış pozitiflerin maliyetinin yüksek olduğu durumlarda önemlidir.

1. **Hatırlama (Recall)**: Gerçek pozitiflerin model tarafından ne kadarının doğru tahmin edildiğini gösterir.

Hatırlama, özellikle yanlış negatiflerin kabul edilemez olduğu durumlarda kritik bir ölçüttür.

1. **F1 Skoru**: Kesinlik ve hatırlama arasında bir denge kurarak modelin genel başarısını ölçen bir metriktir. Özellikle dengesiz veri setlerinde yararlıdır.

F1 skoru, modelin hem kesinliğini hem de hatırlamasını dikkate alarak daha dengeli bir değerlendirme sunar.

1. **Mikro ve Makro Ortalama**: Çok sınıflı sınıflandırma problemlerinde, her bir sınıf için kesinlik, hatırlama ve F1 skorlarının hesaplanmasının yanı sıra, mikro ve makro ortalamalar da alınabilir. Mikro ortalama, tüm sınıfların toplamını dikkate alırken, makro ortalama her bir sınıfın sonuçlarını eşit ağırlıkla değerlendirir.

Bu ölçütler, modelin performansını daha iyi anlamak ve karşılaştırmak için kullanılacaktır. Ayrıca, kategoriler arasındaki dengesizliklerin modeller üzerindeki etkisi de detaylı bir şekilde incelenecektir. Dengesiz veri setleri, bazı sınıfların daha fazla temsil edilmesine neden olabilir ve bu durum, modelin genel performansını olumsuz etkileyebilir. Bu nedenle, sınıflar arası dengeyi sağlamak amacıyla çeşitli stratejiler uygulanacaktır.

Bu çalışma, Türkçe doğal dil işleme alanında önemli bir katkı sunmayı ve sınıflandırma süreçlerini optimize etmeyi hedeflemektedir. Modelin başarısını artırmak için, farklı parametre ayarlamaları, çapraz doğrulama teknikleri ve hiperparametre optimizasyonu gibi yöntemler de değerlendirilecektir. Sonuç olarak, bu süreçlerin her biri, Türkçe haber metinlerinin sınıflandırılmasında daha etkili ve güvenilir bir model elde etmek amacıyla titizlikle uygulanacaktır

**3. SONUÇLAR**

Bu bölümde, Türkçe haber metinlerinin sınıflandırılması için kullanılan modellerin performans sonuçları detaylı bir şekilde sunulacaktır. Modellerin değerlendirilmesi, doğruluk oranı, F1 skoru ve kesinlik (precision) gibi metriklerle gerçekleştirilmiştir. Elde edilen bulgular, her bir modelin etkinliğini ve sınıflandırma sürecinde karşılaşılan zorlukları ortaya koymaktadır.

**3.1.Model.Performansının Değerlendirilmesi**

Aşağıdaki tablo, kullanılan sınıflandırma modellerinin performansını özetlemektedir:

| **Model** | **Doğruluk Oranı** | **F1 Skoru** | **Kesinlik (Precision)** |
| --- | --- | --- | --- |
| **Naive Bayes** | %85 | 0.83 | 0.84 |
| **Destek Vektör Makineleri (SVM)** | %82 | 0.80 | 0.81 |
| **Lojistik Regresyon** | %80 | 0.78 | 0.79 |

**3.1.1. Naive Bayes**

Naive Bayes modeli, %85'lik bir doğruluk oranı ile en yüksek performansı göstermiştir. Bu modelin başarısı, özellikle metin sınıflandırma alanındaki basit ve etkili yapısından kaynaklanmaktadır. Naive Bayes'in avantajları arasında, hızlı öğrenme süreci ve yüksek boyutlu verilerle iyi başa çıkabilme yeteneği bulunmaktadır. Ancak, bu modelin sınırlılığı, bağımsızlık varsayımına dayanmasıdır; bu, kelimelerin birbirinden bağımsız olduğu varsayımını içerir. Bu nedenle, bazı durumlarda bağlamı yeterince iyi yakalayamayabilir.

**3.1.2. Destek Vektör Makineleri (SVM)**

SVM, %82'lik bir doğruluk oranı ile ikinci en iyi performansı göstermiştir. SVM, verileri en iyi ayıran hiper düzlemi bulma yeteneği ile dikkat çeker. Ancak, yüksek boyutlu veri setlerinde eğitim süresi daha uzun olabilir. SVM'nin başarısı, özellikle doğru parametre ayarları ile artırılabilir.Fakat, dengesiz verisetlerinde SVM'nin performansı düşebilir; bu, bazı sınıfların yeterince temsil edilmemesi durumunda daha belirgin hale gelir.

**3.1.3. Lojistik Regresyon**

Lojistik regresyon, %80'lik bir doğruluk oranı ile en düşük performansı sergilemiştir. Ancak, bu modelin yorumlanabilirliği oldukça yüksektir ve sonuçların anlaşılmasını kolaylaştırır. Lojistik regresyon, doğrusal bir model olduğu için, karmaşık ilişkileri yakalamakta zorlanabilir. Bu nedenle, daha karmaşık veri setlerinde diğer modeller kadar etkili olmayabilir.

**3.2. Kategori Tahmini Sonuçları**

Modellerin performansı, kategorilere göre değişiklik göstermiştir. Aşağıda, her bir kategori için elde edilen doğruluk oranları verilmiştir:

| **Kategori** | **Doğruluk Oranı** |
| --- | --- |
| **Ekonomi** | %90 |
| **Spor** | %85 |
| **Sağlık** | %80 |
| **Teknoloji** | %78 |
| **Eğitim** | %65 |

**3.2.1. Ekonomi Kategorisi**

Ekonomi kategorisindeki haberler, %90'lık bir doğruluk oranı ile en yüksek başarıyı göstermiştir. Bu durum, ekonomi haberlerinin genellikle belirli bir dil ve yapı içerdiği gerçeği ile ilişkilidir. Ekonomi haberleri, daha az çeşitlilik ve daha belirgin anahtar kelimeler barındırdığı için, modellerin bu kategoriyi daha iyi öğrenmesini sağlamıştır.

**3.2.2. Spor Kategorisi**

Spor kategorisi, %85'lik bir doğruluk oranı ile ikinci en iyi sonucu elde etmiştir. Spor haberleri de belirli kalıplar ve terminoloji içerdiğinden, modeller bu kategoriyi etkili bir şekilde sınıflandırabilmiştir.

**3.2.3. Sağlık Kategorisi**

Sağlık kategorisindeki haberler, %80'lik bir doğruluk oranı ile başarılı bir şekilde sınıflandırılmıştır. Ancak, sağlık alanındaki haberler bazen karmaşık terimler ve jargona sahip olabilmektedir. Bu durum, modelin öğrenme sürecini zorlaştırabilir.

**3.2.4. Teknoloji Kategorisi**

Teknoloji kategorisi, %78'lik bir doğruluk oranı ile performans açısından orta seviyede kalmıştır. Teknoloji haberleri, sürekli değişen bir alan olduğundan, modelin güncel terimleri ve kavramları öğrenmesi zorlaşabilir.

**3.2.5. Eğitim Kategorisi**

Eğitim kategorisi, %65'lik bir doğruluk oranı ile en düşük başarıyı sergilemiştir. Bu durum, eğitim haberlerinin az temsil edilmesi ve çeşitliliğinin fazla olmasından kaynaklanmaktadır. Az sayıda örnek, modelin bu kategoriyi öğrenmesini zorlaştırmakta ve sonuç olarak doğruluk oranını düşürmektedir.

**3.3. Dengesiz Veri Setinin Etkileri**

Sonuçlar, veri setindeki dengesizliklerin sınıflandırma modelleri üzerindeki etkisini açıkça ortaya koymaktadır. Özellikle az temsil edilen kategorilerde (örneğin Eğitim), modellerin doğruluk oranları düşüş göstermiştir. Bu durum, modelin öğrenme sürecinde yeterli örnek sayısına sahip olmamasının bir sonucudur. Dengesiz veri setleri, bazı sınıfların daha fazla temsil edilmesine neden olabilir ve bu, modelin genel performansını olumsuz etkileyebilir.

**3.4. Gelecek Çalışmalar İçin Öneriler**

Bu çalışmanın sonuçları, sınıflandırma süreçlerini optimize etmek için bazı öneriler sunmaktadır:

1. **Veri Dengeleme**: Az temsil edilen kategoriler için veri artırma teknikleri uygulanabilir. Bu sayede, eğitim verilerinin dengelenmesi sağlanarak modelin bu kategorilerdeki performansı artırılabilir.
2. **Hiperparametre Optimizasyonu**: Modellerin hiperparametre ayarlarının optimize edilmesi, performansın artırılmasına yardımcı olabilir. Özellikle SVM ve Naive Bayes için bu ayarlamalar kritik öneme sahiptir.
3. **Ensemble Yöntemleri**: Farklı modellerin bir araya getirilmesiyle oluşturulan ensemble yöntemleri, genel performansı artırabilir. Örneğin, Naive Bayes ve SVM modellerinin bir arada kullanılması, daha iyi sonuçlar elde edilmesine olanak tanıyabilir.
4. **Gelişmiş Özellik Çıkarımı**: Daha karmaşık özellik çıkarım tekniklerinin (örneğin, kelime gömme yöntemleri) kullanılması, modelin metinleri daha iyi anlamasına yardımcı olabilir.

Sonuç olarak, bu çalışma, Türkçe haber metinlerinin sınıflandırılması alanında önemli bulgular sunmakta ve gelecekteki araştırmalara ışık tutmaktadır. Modellerin performansı, dengesiz veri setlerinin etkisi ve geliştirme önerileri, daha iyi sınıflandırma sonuçları elde etmek için dikkate alınması gereken önemli unsurlardır.

**KAYNAKÇA**

1. Scikit-learn Kullanıcı Belgeleri
2. NLTK Kütüphanesi
3. TF-IDF ve Metin Madenciliği Makaleleri
4. Türkçe Doğal Dil İşleme Kaynakları
5. Kaggle - Turkish News Dataset (<https://www.kaggle.com/datasets/suleymancan/turkishnews70000>)
6. Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing. Pearson.
7. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.