Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 15(1), 57-70, 2025
Journal of the Institute of Science and Technology, 15(1), 57-70, 2025

ISSN: 2146-0574, eISSN: 2536-4618

Bilgisayar Mühendisliği / *Computer Engineering* **DOI: 10.21597/jist.1537627**

Araştırma Makalesi / *Research Article*

Geliş tarihi / *Received*: 24.08.2024 **Kabul tarihi / *Accepted*: 19.10.2024**

Atıf İçin: Veziroğlu, M. ve Bucak, İ. Ö. (2025). Haber Sınıflandırma Sistemlerinde Naive Bayes ve Makine Öğrenmesi Algoritmaları Arasında Performans Karşılaştırması. *Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 15(1), 57-70.

To Cite: Veziroğlu, M. & Bucak, İ. Ö. (2025). Makalenin İngilizce Başlığı. *Journal of the Institute of Science and Technology*, 15(1), 57-70.

Haber Sınıflandırma Sistemlerinde Naive Bayes ve Makine Öğrenmesi Algoritmaları Arasında Performans Karşılaştırması

Merve VEZİROĞLU[1*], İhsan Ömür BUCAK[2]

Öne Çıkanlar:
- Naive Bayes
- Makine öğrenmesi
- Haber sınıflandırması

Anahtar Kelimeler:
- Naive Bayes
- Makine öğrenmesi
- Haber sınıflandırması
- Doğal dil işleme
- Veri ön işleme

ÖZET:

Dijital içerikteki artış, özellikle haber sınıflandırma gibi metin odaklı görevlerde otomatik sınıflandırma yöntemlerine duyulan ihtiyacı büyük ölçüde artırmıştır. Bu noktada Doğal Dil İşleme (DDİ) teknikleri, büyük veri setlerinde insan müdahalesi olmaksızın verimli sonuçlar üretebilme potansiyeline sahiptir. Bu çalışma, haber başlıklarını kategorilere ayırmayı amaçlayan, Python ile geliştirilmiş bir Naive Bayes (NB) tabanlı sınıflandırma sistemini tanıtmaktadır. NB algoritmaları, basitlikleri ve hızlı hesaplama özellikleri nedeniyle metin sınıflandırma problemlerinde öne çıkmaktadır. BBC News başlıklarından oluşan veri kümesi; teknoloji, iş dünyası, spor, eğlence ve siyaset gibi farklı kategorileri kapsamaktadır. Veri ön işleme sürecinde metin temizleme, durdurma kelimelerin çıkarılması ve Sayım Vektörleştirme ile metnin sayısal verilere dönüştürülmesi gibi adımlar yer almıştır. Bu süreç, doğru ve etkili sınıflandırma için kritik bir rol oynamaktadır. Çalışma kapsamında beş farklı NB varyantı incelenmiştir: Gaussian, Multinomial, Complement, Bernoulli ve TAN. Sonuçlar, Multinomial NB'nin %98.53 doğruluk oranıyla en iyi performansı sergilediğini ortaya koymuştur. Complement NB %98.31, TAN %98.20, Bernoulli %96.74, Gaussian NB ise %91.79 ile %92.92 arasında değişen doğruluk oranlarına sahiptir. Bunun yanı sıra NB algoritmaları, Lojistik Regresyon, Rastgele Orman, Doğrusal Destek Vektör Sınıflandırıcısı ve Çok Katmanlı Algılayıcı gibi gelişmiş makine öğrenimi algoritmalarıyla karşılaştırılmıştır. Çok Katmanlı Algılayıcı, %98.31 doğruluk oranı ile öne çıkarken, diğer algoritmalar da %97'nin üzerinde başarı elde etmiştir. Bu çalışma, NB algoritmalarının haber sınıflandırma problemlerinde güçlü, güvenilir ve etkili bir çözüm sunduğunu göstermektedir. Özellikle Multinomial ve Complement NB varyantları, yüksek doğruluk oranları ile dikkat çekmektedir. Gelecekteki araştırmalar, daha geniş veri setleri ve yeni yaklaşımlar ile bu algoritmaların performanslarını daha da geliştirmeyi hedeflemektedir.

Performance Comparison between Naive Bayes and Machine Learning Algorithms for News Classification Systems

Highlights:
- Naive Bayes
- Machine learning
- News classification

Keywords:
- Naive Bayes
- Machine learning
- News classification
- Natural language processing
- Data preprocessing

ABSTRACT:

The rapid increase in digital content, particularly in text-based tasks like news classification, has significantly amplified the demand for automated classification methods. At this point, Natural Language Processing (NLP) techniques offer the potential to efficiently generate results from large datasets without human intervention. This study presents a Naive Bayes (NB)-based classification system, developed using Python, aimed at categorizing news headlines. NB algorithms are favored for text classification problems due to their simplicity and fast computation. The dataset used, derived from BBC News headlines, covers diverse categories such as technology, business, sports, entertainment, and politics. The data preprocessing phase included steps such as text cleaning, removing stop words, and converting the text into numerical data using Count Vectorization. This process plays a critical role in ensuring accurate and effective classification. Five different NB variants were examined in this study: Gaussian, Multinomial, Complement, Bernoulli, and Tree-Augmented Naive Bayes (TAN). The results showed that Multinomial NB delivered the best performance with an accuracy rate of 98.53%. Complement NB achieved 98.31%, TAN 98.20%, Bernoulli 96.74%, while Gaussian NB ranged between 91.79% and 92.92%. Additionally, NB algorithms were compared with advanced machine learning algorithms such as Logistic Regression, Random Forest, Linear Support Vector Classifier, and Multi-Layer Perceptron. The Multi-Layer Perceptron stood out with an accuracy rate of 98.31%, while the other algorithms also surpassed 97% accuracy. This study demonstrates that NB algorithms provide a robust, reliable, and effective solution for news classification problems, with the Multinomial and Complement variants showing particularly high accuracy. Future research will aim to further enhance the performance of these algorithms using larger datasets and new approaches.

[1] Merve VEZİROĞLU (**Orcid ID:** 0000-0002-4428-1188), [2]İhsan Ömür BUCAK (**Orcid ID:** 0000-0002-9112-3932), Iğdır Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Iğdır, Türkiye

***Sorumlu Yazar/*Corresponding Author**: Merve VEZİROĞLU, e-mail: mervearg@gmail.com

GİRİŞ

Dijital içeriğin hızlı artışı, çevrimiçi erişilebilen büyük miktarda verilerin ortaya çıkmasına yol açmıştır. Sonuç olarak, metinsel verileri otomatik olarak sınıflandırmak ve kategorize etmek için verimli yöntemlere olan talep giderek artmaktadır. Doğal Dil İşleme (DDİ) uygulamalarının önemli bir parçası olan haber sınıflandırması, haber başlıklarını içeriklerine dayalı olarak önceden tanımlanmış belirli kategorilere otomatik olarak atamayı hedefler. Bu görev, içerik önerisi, bilgi erişimi ve duygu analizi gibi önemli pratik sonuçlara haizdir.

Bu çalışmada, Python programlama dili kullanılarak gerçekleştirilen kapsamlı bir haber sınıflandırma görevi açıklanmaktadır. Özellikle haber başlıklarını çeşitli sınıflara doğru bir şekilde kategorize etmek için Naive Bayes algoritmalarının gücünü kullanma üzerine odaklanmış bulunmaktayız. Metin sınıflandırma görevlerinde basitliği ve etkinliğiyle bilinen, iyi yapılandırılmış ve yaygın olarak kullanılan bir makine öğrenmesi algoritması olan Naive Bayes, bu araştırmanın temelini oluşturmaktadır.

Ana hedeflerimizden ilki, sağlam bir haber sınıflandırma sistemi oluşturmanın karmaşıklıklarını incelemek, ikincisi ise farklı Naive Bayes algoritmalarının performansını titizlikle değerlendirmektir. Bu çalışmada, teknoloji, iş, spor, eğlence ve politika olmak üzere beş ana kategoriye ayrılmış haber başlıkları, saygın bir kuruluş olan BBC News Corpus'tan (2006) temin edilen bir veri seti üzerinde inşa edilmiştir. Ayrıntılı bir analiz yoluyla, kategorilerin dağılımına, başlıkların özelliklerine ve iç yapılarına dair derinlemesine bir anlayış kazandırmayı amaçlamaktayız. Özellikle, veri temizleme, durdurma kelimelerinin kaldırılması ve sayım vektörleştirme tekniği kullanarak özellik çıkarımı gibi kapsamlı veri ön işleme yöntemlerini uygulamaktayız (Patel ve Meehan, 2021). Veri seti eğitim ve test setlerine bölünmekte, dönüştürülen sayısal vektörler ise Naive Bayes algoritmalarını eğitmek amacıyla giriş verisi olarak kullanılmaktadır (Saritas ve Yasar, 2019).

Araştırmamız, Gaussian Naive Bayes, Multinomial Naive Bayes, Complement Naive Bayes ve Bernoulli Naive Bayes olmak üzere beş farklı Naive Bayes varyantı üzerine odaklanmaktadır (Chen, Webb ve ark., 2020). Her algoritma, doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1 skoru (F1 score) gibi iyi bilinen performans ölçütleri kullanılarak test verileri üzerinde titizlikle eğitilmekte ve değerlendirilmektedir (Powers, 2020). Ayrıca, Naive Bayes algoritmalarının performansı, Lojistik Regresyon, Rastgele Orman, Doğrusal Destek Vektör Makinesi (DVS), Çizgi Küme Analizi (ÇKA) Sınıflandırıcısı, Karar Ağacı ve K-En Yakın Komşu (K-EYK) gibi diğer iyi bilinen makine öğrenmesi sınıflandırıcıları ile kapsamlı bir şekilde karşılaştırılmaktadır (Mahesh, 2020). Sonuçların ayrıntılı analizi, haber sınıflandırması bağlamında her algoritmanın göreceli güçlü ve zayıf yönlerini vurgulamak amacıyla gerçekleştirilmektedir.

Bulgularımız, tüm Naive Bayes algoritmalarının haber sınıflandırması görevlerinde hatırı sayılır bir doğruluk sergilediğini ve özellikle Multinomial Naive Bayes ve Complement Naive Bayes'in etkili varyantlar olduğunu açıkça ortaya koymaktadır. Ayrıca, ÇKA Sınıflandırıcısı, haber kategorize etmede en iyi performans gösteren makine öğrenmesi sınıflandırıcısı olarak öne çıkmaktadır.

Sonuç olarak, bu çalışma, haber sınıflandırmasında Naive Bayes algoritmalarının etkili bir uygulamasını içermekte olup, kapsamlı bir veri ön işleme sürecinin ve titiz bir algoritma seçiminin önemini vurgulamaktadır. Ayrıca, bu araştırmanın kapsamlı sonuçlarının ve sağladığı farkındalıkların, doğal dil işleme alanında çalışan araştırmacılar ve uygulayıcılar için değerli bir kaynak sunacağı ve verimli haber kategorize etme veya sınıflandırma sistemlerinin geliştirilmesine önemli katkılarda bulunacağı kanaatindeyiz.

Haber sınıflandırması, haber makalelerini etkili bir şekilde kategorize etmek amacıyla çeşitli yaklaşım ve algoritmaların araştırıldığı önemli bir araştırma alanı olarak bilinmektedir. Çok sayıda çalışma, Naive Bayes, Destek Vektör Makinaları (Support Vector Machines (SVM)), Rastgele Orman, Lojistik Regresyon ve Yapay Sinir Ağları (Artificial Neural Networks) gibi çeşitli makine öğrenmesi ve derin öğrenme algoritmalarının etkinliğini araştırmıştır.

Naive Bayes algoritması, basitliği, verimliliği ve etkinliği nedeniyle haber sınıflandırmasında özellikle tercih edilmektedir. Son araştırmalar, farklı bağlamlarda özellikle bu algoritmanın performansını artırmaya yönelik bir odaklanma içerisindedir.

Rana ve arkadaşları, başlıklara dayalı bir haber sınıflandırmasını araştırmış ve Naive Bayes'in etkinliğini vurgulamışlardır. Özellik seçimi, ön işleme teknikleri ve sınıflandırma yaklaşımlarını inceleyerek, Naive Bayes'in bu görev için basit ama güçlü bir tercih olduğunu sonucuna varmışlardır (Rana, Khalid ve ark., 2014) .

Shahi ve Pant, Nepal haber makalelerini sınıflandırmak üzere Naive Bayes'i DVM ve Yapay Sinir Ağları ile karşılaştırmışlar ve de TF-IDF vektörizasyonu kullanmışlardır. Naive Bayes, üstün doğruluğu ve gösterdiği F1 skor performansı ile Nepal haber sınıflandırması için uygun bir seçim olduğu kanaatini oluşturmuştur (Shahi ve Pant 2018) .

Chy ve arkadaşları, Bangla haber makalelerini kategorize etmek üzere Naive Bayes sınıflandırıcısını uygulamışlar, kök bulma, durdurma kelimelerini kaldırma ve TF-IDF vektörizasyonu kullanmışlardır. Çalışmalarında, karşılaştırılan algoritmalar arasında Naive Bayes'in en yüksek doğruluğa sahip olduğuna vurgu yapmışlar ve sınıflandırma sonuçlarını iyileştirmek amacıyla ön işleme tekniklerini önermişlerdir (Chy, Seddiqui ve ark., 2014).

Bracewell ve arkadaşları, Japonca ve İngilizce makaleleri kategorize etmek üzere Naive Bayes algoritmasını kullanarak çapraz dilli haber sınıflandırmasını araştırmışlardır. Yaklaşımları, kategori sınıflandırmasını konu keşfi ile birleştirerek, Naive Bayes'in diller arasında üstün doğruluğu ve F1 skoru ile tercih edildiğini göstermektedir (Bracewell, Yan ve ark., 2009).

Albahr ve Albahar, haber tespiti için çeşitli makine öğrenmesi sınıflandırıcılarını değerlendirmişler, Naive Bayes ile en iyi sonuçları elde etmişlerdir. Çalışmaları, Naive Bayes'in sınıflandırma görevlerindeki dayanıklılığına vurgu yapmış bulunmaktadır (Albahr ve Albahar 2020).

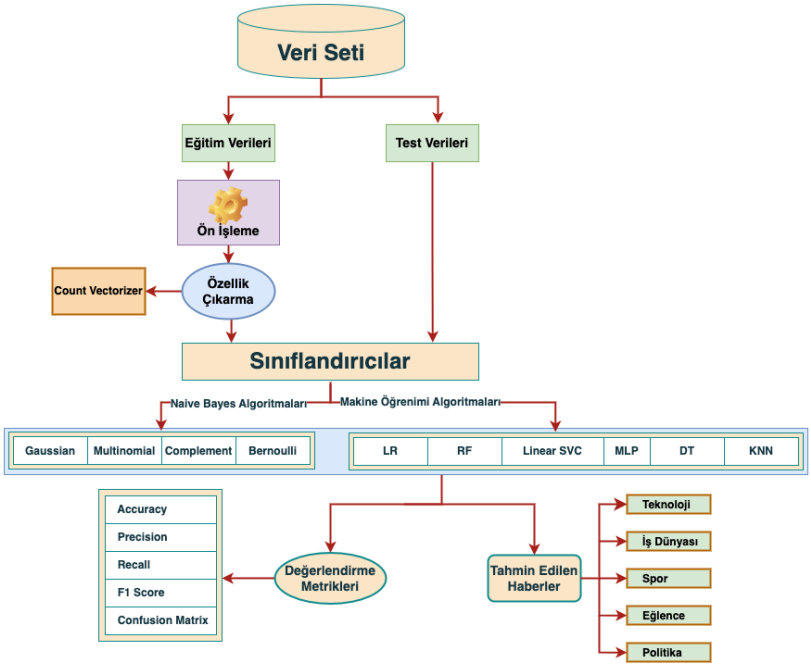
Sristy ve Somayajulu, etiketli ve etiketlenmemiş veri setlerinden yararlanarak haber makalesi sınıflandırması için Naive Bayes ile yarı denetimli bir yaklaşım önermişlerdir. Naive Bayes, diğer yarı denetimli algoritmalardan daha iyi performans göstermiş, özellikle sınırlı etiketli veri ile doğruluğu artırma potansiyeli taşıdığına vurgu yapılmıştır (Sristy ve Somayajulu 2012).

Granik ve Mesyurar, Facebook haber gönderilerini kullanarak haber tespiti için Naive Bayes tabanlı bir sistem geliştirmişler ve %74 sınıflandırma doğruluğu elde etmişlerdir. Çalışmaları, Naive Bayes'in gerçek dünya içerik sınıflandırmasına ilişkin pratik bir uygulamayı içermektedir (Granik ve Mesyura 2017).

Bu çalışmalar, Naive Bayes algoritmalarının haber sınıflandırmasındaki çok yönlülüğünü ve etkinliğini kapsamlı bir şekilde vurgulamaktadır. Özellik seçimi, ön işleme ve algoritma entegrasyonu için farklı stratejiler, sınıflandırma doğruluğunu optimize etmeye yönelik değerli kavrayışlar ve öngörüler sunmaktadır. Ayrıca, Naive Bayes algoritmasını tamamlayıcı tekniklerle birleştirmenin, haber sınıflandırma sistemlerinin daha da iyileşmesine katkıda bulunacağına dair önerilerde bulunmaktadırlar.

MATERYAL VE METOT

Bu bölümde, haber sınıflandırmasını gerçekleştirmek ve Naive Bayes algoritmalarının etkinliğini değerlendirmek amacıyla gerekli olan araştırma yöntembilimi ayrıntılı bir şekilde ele alınacaktır. Burada yöntembilimi, veri toplama, veri ön işleme, model eğitimi ve performans değerlendirmesi safhalarına ilişkin yapılandırılmış bir süreci içermektedir. Aşağıda, gerçekleştirilen çalışmanın titizliğini vurgulamak ve geçerliliğini sağlamak amacıyla uygulanan temel yöntemler özetlenmiştir. Yöntemsel adımlar ise Şekil 1'de ayrıntılı olarak sunulmaktadır.



Şekil 1. Genel Çalışma Yordamı

Veri Toplama

Çalışmamızda kullanılan veriler, çeşitli konulara göre kategorize edilmiş haber başlıklarından oluşan BBC News Corpus (2006) veri setinden elde edilmiştir. Veri seti, her haber öğesinin ait olduğu konuyu belirten kategori (category) kolonu ile haber öğesinin gerçek içeriğini gösteren metin (text) kolonundan oluşan iki ana başlıktan ibarettir.

Veri seti, teknoloji, iş, spor, eğlence ve politika olmak üzere beş farklı kategoriye ayrılmış haber başlıklarını içermektedir. Her kategoriye ait veri noktalarının dağılımı aşağıda Çizelge 1'de ayrıntılı olarak sunulmuştur.

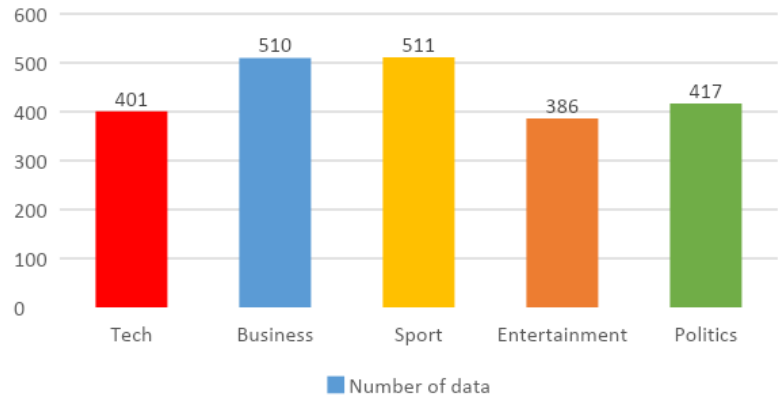
Çizelge 1. Veri Setinin Dağılımı

Kategori	Veri sayısı
Teknoloji	401
İş Dünyası	510
Spor	511
Eğlence	386
Politika	417
Toplam	**2225**

Bu veri seti, haber sınıflandırması için NB algoritmalarının kapsamlı bir şekilde araştırılmasına olanak tanıyan çeşitli haber başlıklarından oluşmaktadır. Veri toplama süreci, farklı kategorilerdeki haber başlıklarının dengeli bir şekilde temsil edilmesini sağlayarak bu araştırmadan güvenilir ve

anlamlı bulgular elde edilmesini kolaylaştırmaktadır. Şekil 2'de, veri setinin detaylarını içeren görsel bir temsil sunulmaktadır.

Veri kümesi, beş farklı kategoriye ait haber başlıklarından oluşmaktadır. Bu kategoriler, Teknoloji, İş Dünyası, Spor, Eğlence ve Politika olarak tasnif edilmiştir. Her bir kategoride mevcut olan veri nokta sayısı, daha önce bahsedildiği üzere Çizelge 1'de ifade edilmiştir.



Şekil 2. Her Bir Sınıf İçin Veri Sayısı Dağılımları

Veri ön işleme

Doğal Dil İşleme (DDİ) alanında ham metin verilerini analiz ve sınıflandırma görevleri için hazır hale getirmek, önemli ön işleme adımlarını gerektirmektedir. Bu çalışmada, BBC News Corpus'tan elde edilen haber başlıkları veri seti, Şekil 3'te gösterildiği üzere çeşitli ön işleme adımlarından geçirilmiştir. Veri ön işlemenin asıl amacı, ham metin verilerini makine öğrenmesi algoritmaları yardımıyla yapılacak daha sonraki analizler için uygun bir yapısal formata dönüştürmektir.

Şekil 3. Model Ön İşleme Teknikleri

Veri temizleme

Veri temizleme aşamasının temel görevi, haber başlıklarındaki noktalama işaretleri ve özel karakterleri kaldırmaktır. Bu unsurlar, metin bağlamına katkıda bulunmadığı gibi analiz sırasında gürültü oluşturma potansiyeline sahiptirler. Düzenli ifadeler kullanılarak noktalama işaretleri ve özel karakterler sistematik bir şekilde ortadan kaldırılmış, böylece yalnızca gerekli metin içerikleri korunmuştur. Ayrıca, bazı haber başlıkları tarihler, istatistikler veya diğer sayısal veriler gibi sayısal değerler içerebilmektedir. Ancak, bu sayısal değerler genellikle sınıflandırma görevleri için alakasız olabilmekte ve model performansını olumsuz yönde etkileyebilmektedir. Bu nedenle, yalnızca metin içeriğine odaklanılmış ve alakasız bilgilerin etkisini azaltmak amacıyla başlıklardan sayısal değerlerin kaldırılması yoluna gidilmiştir. Tüm bu adımlar, işlenmiş verilerin makine öğrenmesi algoritmaları

aracılığıyla etkin bir şekilde sınıflandırılabilmesi için gerekli metin içeriğinin doğru bir şekilde yansıtılmasını sağlamak amacıyla uygulanmıştır.

Durdurma kelimelerini kaldırma

Durdurma kelimeleri, gerek makaleler ve gerekse yaygın dil kullanımında sırasıyla ("the," "a," "an"), ("is," "are," "in") gibi örnek kelimeler şeklinde karşımıza çıkan ve sık kullanılan ancak önemli bir anlam yüklenilemeyecek kelimeler bütünü olarak tanımlanabilir. Bu çalışmada analizi daha verimli hale getirmek ve korunan kelimelerin ilgi düzeylerini artırmak üzere durdurma kelimelerinin kaldırılma sürecini kritik bir işlem adımı olarak uygulamış bulunuyoruz. Önceden tanımlanan İngilizce durdurma kelimeleri listesi kullanılarak, bu kelimeler haber başlıkları veri setinden sistematik olarak çıkarılmıştır. Bu yaklaşım, dikkate alınan kelime sayısını etkili bir şekilde azaltarak sonraki analizleri basitleştirmiş ve daha bilgilendirici özelliklere odaklanma imkânı sağlamıştır.

Özellik çıkarımı

Tokenizasyon (Tokenization) bir ürün veya hizmet verilerinin farklı değerlere dönüştürülerek daha güvenli hale gelmesini ifade eder. Tokenizasyon, metni bireysel kelimelere veya *token* adı verilen daha küçük varlıklara ayırmayı içerir ve haber başlıklarını anlamlı birimlere bölerek kelime düzeyinde daha ileri analiz yapmayı mümkün kılar. İşte bu çalışmada haber başlıklarında bulunan özgün kelimelerin oluşturduğu bir kelime dağarcığı listesi inşa edilmesinde bu *tokenlaştırılmış* kelimelerden yararlanılmıştır. Bu kelime listesi, veri seti boyunca tutarlı kelime temsili sağlayarak özellik çıkarımı için bir temel oluşturmuştur. Ardından, her bir belirli haber başlığı için her kelimenin ne sıklıkta göründüğünü ölçen bir *terim sıklığı* (term frequency (TF)) değeri hesaplanmıştır. Bu sıklık verisi, metin verilerini makine öğrenmesi algoritmaları için uygun sayısal vektörlere dönüştürmek açısından son derece önemlidir. Zira bu veri, her satırın bir haber başlığını ve her kolonun ise kelime dağarcığı içindeki özgün bir kelimeyi ifade ettiği özel bir matrisin oluşturulmasını sağlamıştır. Ayrıca matris girişleri, her kelimenin ilgili başlıkta ne sıklıkta göründüğünü yansıtmaktadır.

Sayım vektörleştirici (count vectorizer)

Sayım vektörleştirme tekniği, metin verilerini sayısal matris temsiline dönüştürmek için kullanılan temel bir yöntemdir. Bu teknik, *tokenize edilmiş* kelimelerden bir kelime dağarcığı oluşturur ve her kelimeye eşsiz bir tam sayı atar. Bu kelime dağarcığını kullanarak, her satırın bir haber başlığına ve her kolonun da kelime dağarcığından özgün bir kelimeye karşı düştüğü bir *doküman terim matrisi* oluşturulmuştur. Matris girişleri, her kelimenin ilgili başlıkta ne sıklıkta göründüğünü gösterirler. Uygulanan veri ön işleme adımları, hem veri setinin kalitesinin artırılmasında hem de makine öğrenmesi algoritmalarıyla gerçekleştirilecek analize hazırlık aşamasında önemli bir rol oynamıştır. Gürültünün giderilmesi, metnin standart hale getirilmesi, hatanın düzeltilmesi ve eksik verilerin işlenmesi yoluyla temizlenen veri seti, doğru ve güvenilir haber sınıflandırması için sağlam bir temel teşkil etmektedir.

Veri temizleme

Makine öğrenmesi (mö) algoritmaları

Makine Öğrenmesi (Machine Learning - MÖ), bilgisayar sistemlerinin verilerden öğrenmesini ve bu verilere dayanarak kararlar veya tahminler yapmasını sağlayan algoritmalar ile istatistiksel modelleri inceleyen bir Yapay Zekâ alt dalıdır. Veri setlerinden modelleri oluşturmak için tasarlanan bu tür algoritmalar yeni verilerdeki örüntüleri tanımlayabilir veya tahminlerde bulunabilir. Makine öğrenmesinin esas amacı, ana özellikleri ve örüntüleri yakalayarak verilerden genelle(ştir)me

yapmaktır. Bu genelle(ştir)me, MÖ algoritmalarının yeni ve önceden görülmemiş verilerle karşılaştıklarında bilgilendirilmiş kararlar veya tahminler yapmalarına olanak sağlar. Bu çalışmada, önceden işlenmiş verileri kullanarak haber sınıflandırması yapmak amacıyla çeşitli MÖ sınıflandırıcıları kullanılmıştır. Kullanılan sınıflandırıcılar, metin sınıflandırma görevlerinde basitlikleri ve etkinlikleri ile bilinen NB algoritmalarının yanı sıra diğer iyi bilinen MÖ algoritmalarını da içermektedir. Bu sınıflandırıcılar, teknoloji, iş, spor, eğlence ve politika gibi önceden tanımlanmış haber başlıklarını ayırt etmek ve sınıflandırma yapmak üzere ön işlemeden geçmiş veri setleri üzerinde eğitilmişlerdir.

Naive bayes (NB) algoritmaları

Naive Bayes, her sınıf içindeki özelliklerin birbirinden bağımsız olduğunu varsayarak Bayes teoremini kullanan olasılıksal bir sınıflandırma yöntemidir. Naive Bayes'in matematiksel süreci şu şekilde açıklanabilir:

$$P(C|X) = \frac{P(X \mid C) * P(C)}{P(X)} \quad (1)$$

Burada, P(C|X), X gözlemlenen kanıtı göz önüne alındığında C sınıfının sonsal olasılığını temsil eder. P(X|C), C sınıfı verildiğinde X kanıtını gözlemleme olasılığıdır. P(C), C sınıfının önsel olasılığıdır ve P(X) kanıt olasılığıdır.

BULGULAR VE TARTIŞMA

Değerlendirme Metrikleri

Gerek makine öğrenmesi algoritmalarının etkinliğini değerlendirmeleri gerekse bilinçli kararlar almaları, performans ölçütlerini hayati derecede önemli kılmaktadır. Bu ölçütler, algoritma performansını değerlendirmek, eniyilemek (optimization) sürecini yönlendirmek, sonuçları raporlamak, hataları ve önyargıları belirlemek, kıyaslama noktaları oluşturmak ve aşırı öğrenmeyi tespit etmek gibi çeşitli görevlerde önemli bir rol oynamaktadır. Bu çalışmada, özellikle sınıflandırma algoritmalarının performansını değerlendirmek ve karşılaştırmak amacıyla yaygın olarak kullanılan ölçütlere yer verilmiştir.

Makine öğrenmesi için önemli performans ölçütleri arasında doğruluk, kesinlik, duyarlılık ve F1 skoru sayılabilir. Doğruluk, doğru tahminlerin toplam tahmin sayısına oranını temsil eder ve modelin genel performansını gösterir. Kesinlik, doğru pozitif olarak tahmin edilen örneklerin, pozitif olarak tahmin edilen toplam örnek sayısına oranıdır. Modelin pozitif tahminlerinin doğruluğunu gösterir. Duyarlılık, modelin doğru biçimde tespit ettiği gerçek pozitif örneklerin oranını ölçer. Modelin doğru pozitif örnekleri tanımlama yeteneğini gösterir. F1 skoru ise kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır. Bu ölçüt, özellikle veri kümesi dengesiz olduğunda kesinlik ve duyarlılık arasında bir denge sağlamaktadır.

$$Accuracy\ (Doğruluk) = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision\ (Kesinlik) = \frac{TP}{TP + FP} \quad (3)$$

$$Recall\ (Duyarlılık) = \frac{TP}{TP + FN} \quad (4)$$

$$F1\ Skoru = 2.\frac{Precision\ .\ Recall}{Precision + Recall} \quad (5)$$

Burada, TP (True Positive), doğru biçimde olumlu olarak tahmin edilen örneklerin sayısını; TN (True Negative), doğru biçimde olumsuz olarak tahmin edilen örneklerin sayısını; FP (False Positive), yanlış biçimde olumlu olarak tahmin edilen örneklerin sayısını; FN (False Negative) ise yanlış biçimde olumsuz olarak tahmin edilen örneklerin sayısını temsil eder.

Çalışmamızda, bu performans ölçütleri aracılığıyla Naive Bayes algoritmaları ve diğer makine öğrenmesi sınıflandırıcılarının etkinliklerinin yanı sıra, sınıflandırma yeteneklerinin de kapsamlı bir değerlendirmesi yapılmıştır.

Deneysel sonuçlar

Bulgular

Farklı NB algoritmalarına ilişkin performans değerlendirmesi sonuçları aşağıda Çizelge 2'de ayrıntılı bir biçimde sunulmuştur. Her bir varyantın doğruluk, kesinlik, duyarlılık ve F1 skoru titizlikle hesaplanmıştır.

Epsilon değeri 0.01 olan Multinomial Naive Bayes algoritması, Naive Bayes varyantları arasında en yüksek performans değerlerine ulaşmıştır: %98.53 doğruluk, %98.55 kesinlik, %98.53 duyarlılık ve %98.54 F1 skoru. Bu sonuçlar, algoritmanın sağlamlığını ve haber sınıflandırma görevleri için uygunluğunu açıkça ortaya koymaktadır.

Farklı parametrelere sahip Multinomial Naive Bayes algoritması da rekabetçi sonuçlar üretmiştir. Multinomial (Alpha, 0.5) %98.53 doğruluk, %98.53 kesinlik, %98.53 duyarlılık ve %98.53 F1 skoru ile hemen hemen benzer performans değerlerine ulaşmıştır. Bu tutarlılık, algoritmanın farklı parametre değerlerinde güvenilirliğini pekiştirmektedir.

Complement Naive Bayes ise %98.31 doğruluk ve benzer kesinlik, duyarlılık ve F1 skorları ile önemli ölçüde iyi sayılabilecek bir performans sergilemiştir. Özellikle dengesiz veri setlerinde etkinliğini hissettirmiştir. Ayrıca, *fit prior* değeri *Yanlış* olarak ayarlanan Bernoulli Naive Bayes algoritması tüm metriklerde %96.74'e ulaşarak, belirli özellik dağılımlarına sahip senaryolarda uygulanabilirliğini kanıtlamıştır.

Bununla beraber, Gaussian Naive Bayes algoritması, %92.92 doğruluk, %93.08 kesinlik, %92.92 duyarlılık ve %92.93 F1 skoru ile nispeten daha düşük performans metrik değerleri ile neticelenmiştir. Bu sonuçlar, Gaussian Naive Bayes'in özelliklerin normal dağılım göstermediği metin sınıflandırma görevlerine daha az uygunluk göstereceği anlamına gelmektedir.

Çizelge 2. Naive Bayes Algoritmasının Sonuçları

Naive Bayes Algoritmaları	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 skoru (F1 score)
Gaussian	92.92%	93.08%	92.92%	92.93%
Gaussian (Var Smoothing)	91.79%	91.79%	91.79%	91.79%
Gaussian (Var Smoothing, 2e-9)	91.79%	91.79%	91.79%	91.79%
Multinomial	98.20%	98.20%	98.20%	98.20%
Multinomial (Fit Prior, False)	98.31%	98.31%	98.31%	98.31%
Multinomial (Alpha, 0.5)	**98.53%**	**98.53%**	**98.53%**	**98.53%**
Multinomial (epsilon = 0.01)	**98.53%**	**98.55%**	**98.53%**	**98.54%**
Multinomial(epsilon = 1e-9)	96.85%	96.91%	96.85%	96.85%
Multinomial (Alpha, 1.5)	98.20%	98.20%	98.20%	98.20%
Complement	98.31%	98.31%	98.31%	98.31%
Complement (Alpha, 0.5)	98.20%	98.20%	98.20%	98.20%
Bernoulli	96.67%	97.01%	96.74%	96.77%
Bernoulli (Fit Prior, False)	96.74%	96.74%	96.74%	96.74%
TAN (Tree-Augmented Naive Bayes) (epsilon = 0.01)	98.20%	98.23%	98.20%	98.20%

Çizelge 3 ise yaygın olarak kullanılan diğer makine öğrenmesi algoritmalarının performans metriklerini sunmaktadır. ÇKA Sınıflandırıcısı %98.31 doğruluk, %98.32 kesinlik, %98.31 duyarlılık ve %98.31 F1 skoru ile en yüksek doğruluk değerlerine ulaşmıştır. Bu değerler, sınıflandırıcının karmaşık örüntüleri ve yüksek boyutlu verileri işleme yeteneğine vurgu yapmaktadır.

Doğrusal Destek Vektör Sınıflandırıcısı (DVS), %97.97 doğruluk, %97.98 kesinlik, %97.97 duyarlılık ve %97.98 F1 skoru ile bu sınıflandırıcıyı yakından takip etmektedir ve haliyle güçlü bir genelle(ştir)me yeteneği sergilemektedir. Ayrıca Rastgele Orman, %97.86 doğruluk ve bilahare takip eden kesinlik, duyarlılık ve F1 skorları ile iyi bir performans göstererek topluluk öğrenmesindeki sağlamlığını kanıtlamıştır.

Lojistik Regresyon, %97.52 doğruluk, ve diğer kesinlik, duyarlılık ve F1 skor değerleri ile rekabetçi sayılabilecek performans değerlerine sahiptir; bu haliyle, metin sınıflandırması için güvenilir ve yorumlanabilir bir model olduğu izlenimi vermektedir. Daha az etkili uçta, örneğin Karar Ağacı algoritması %82.69 doğruluk değeri ile karmaşık veri setlerinde aşırı uyum eğiliminin bir zaafı olarak nispeten düşük performans sergilemiştir. K-EYK, %55.84 doğruluk ile en düşük performans değerini sergilemiştir; haliyle bu da büyük ve karmaşık veri setleriyle başa çıkmadaki sınırlamalarını açığa vurmaktadır.

Çizelge 3. Diğer İyi Bilinen MÖ Algoritmaları İçin Sonuçlar

MÖ Algoritmaları	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 skoru (F1 score)
Lojistik Regresyon	97.52%	97.52%	97.52%	97.52%
Rastgele Orman	97.86%	97.91%	97.86%	97.87%
Doğrusal DVS	97.97%	97.98%	97.97%	97.98%
Radyal Bazlı Fonksiyon DVM Radial Basis Function (RBF) DVM	91.57%	91.57%	91.57%	91.57%
Doğrusal Ayırtaç Analizi	38.20%	38.20%	38.20%	38.20%
ÇKA Sınıflandırıcısı	***98.31%***	***98.32%***	***98.31%***	***98.31%***
Karar Ağacı	82.69%	82.84%	82.69%	82.72%
KNN	55.84%	80.88%	55.84%	56.59%
Ekstra Ağaçlar (Extra Trees)	95.61%	95.61%	95.61%	95.61%
Gradyan Hızlandırma (Gradient Boosting)	94.04%	94.04%	94.04%	94.04%
Pasif Saldırgan Sınıflandırıcı (Passive Agressive Classifier)	96.62%	96.62%	96.62%	96.62%
AdaBoost	65.73%	65.73%	65.73%	65.73%
Torbalama Sınıflandırıcısı (Bagging Classifier)	82.35%	82.35%	82.35%	82.35%
Sırt Sınıflandırıcı (Ridge Classifier)	96.29%	96.29%	96.29%	96.29%
Algılayıcı (Perceptron)	96.17%	96.17%	96.17%	96.17%

NB algoritmaları diğer makine öğrenmesi algoritmaları ile karşılaştırıldıklarında, her iki grubun veri setinin doğasına bağlı olarak son derece yüksek performans gösterebilecekleri açıkça söylenebilir. Multinomial ve Complement Naive Bayes gibi NB varyantları, ayrık ve dengesiz veri setlerini etkili bir şekilde işleme yetenekleri sayesinde son derece yüksek performans göstermiştir.

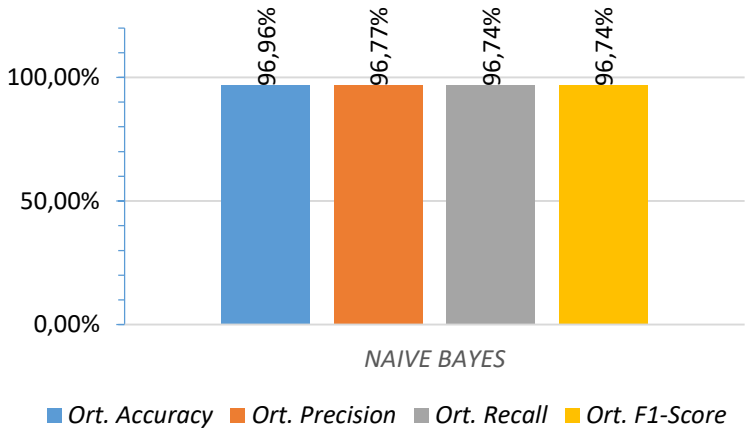
Buna karşılık, ÇKA Sınıflandırıcısının üstün performansı, derin öğrenme modellerinin metin verilerindeki mevcut karmaşık örüntüleri yakalama gücüne vurgu yapmaktadır. Farklı haber başlıkları kategorilerinde iyi genelle(ştir)me yapabilme yeteneği, onu yüksek doğruluk gerektiren görevler için optimal bir seçim haline getirmektedir.

Kapsamlı performans değerlendirmesi, epsilon değeri 0.01 olan Multinomial, Complement Naive Bayes ve ÇKA Sınıflandırıcısı'nın doğru haber sınıflandırması için en iyi adaylar arasında olduğunu açıkça göstermektedir. Veri setinin spesifik gereksinimleri ve özelliklerine bağlı olarak, bu algoritmaların her biri metin sınıflandırma görevlerinde tercih edilen bir seçenek olabilir. Algoritma

seçiminin, veri seti özelliklerine ve uygulama ihtiyaçlarına dayalı olarak yapılmasının önemi vurgulanmaktadır.

Sonuç olarak, NB algoritmaları belirli türdeki metin verileri için son derece etkili olmasına karşılık, ÇKA Sınıflandırıcısı genel tutarlılığı ve doğruluğu ile öne çıkarak, haber sınıflandırmasında pratik uygulamalar için sağlam bir çözüm olarak durmaktadır.

Şekil 4, NB algoritmalarının ortalama performansını göstermektedir. Bu sonuçlar, Naive Bayes algoritmalarının haber sınıflandırmasında oldukça yüksek bir performans sergilediğini göstermektedir. Özellikle Multinomial Naive Bayes ve varyasyonları en yüksek performansı sergilemişlerdir. Multinomial Naive Bayes (Alpha, 0.5) %98.53 doğruluk, %98.53 kesinlik, %98.53 duyarlılık ve %98.53 F1 skoru ile en yüksek performans sonuçlarını üretmiştir. Gaussian Naive Bayes ise genellikle diğer varyasyonlara göre biraz daha düşük performans sergilemiştir. Ancak yine de yüksek doğruluk oranlarına sahiptir.



Şekil 4. NB Algoritmalarının Ortalama Performans Değerleri

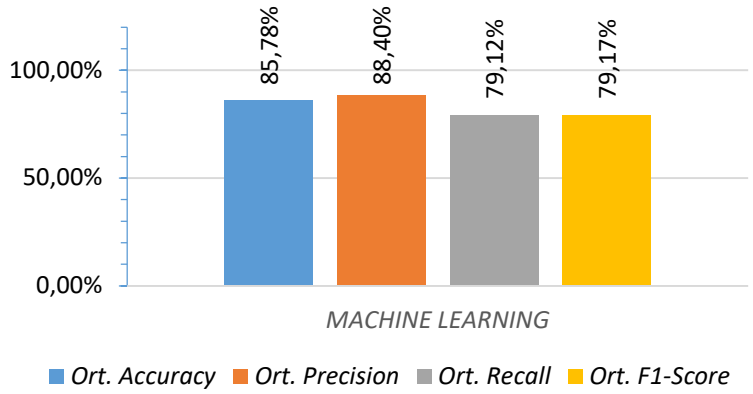
Şekil 5, Lojistik Regresyon, Rastgele Orman, Doğrusal DVS, RBF DVM, Doğrusal Ayırtaç Analizi, ÇKA Sınıflandırıcısı, Karar Ağacı, K-EYK, Ekstra Ağaçlar, Gradyan Hızlandırma, Pasif Saldırgan Sınıflandırıcısı, AdaBoost, Torbalama Sınıflandırıcısı, Sırt Sınıflandırıcısı ve Algılayıcı algoritmalarının ortalama performanslarını göstermektedir. Bu sonuçlar, diğer makine öğrenmesi algoritmalarının NB algoritmalarına kıyasla daha düşük performans sergilediklerine işaret etmektedir. Doğrusal Ayırtaç Analizi ve K-EYK algoritmaları en düşük performansı göstermişlerdir. Özellikle Doğrusal Ayırtaç Analizi, %38.20 doğruluk yüzdesi ile oldukça düşük bir performans sergilemiştir. Bunun aksine, Lojistik Regresyon, Rastgele Orman, Doğrusal DVS ve ÇKA Sınıflandırıcısı algoritmaları oldukça yüksek performans göstermişlerdir. Özellikle, ÇKA Sınıflandırıcısı %98.31 doğruluk, %98.32 kesinlik, %98.31 duyarlılık ve %98.31 F1 skoru ile en yüksek performans sergileyen algoritma olmuştur.

NB algoritmaları, haber sınıflandırması problemlerinde diğer makine öğrenmesi algoritmalarına göre daha yüksek performans sergilemişlerdir. Ortalama doğruluk, kesinlik, duyarlılık ve F1 skoru metriklerinde NB algoritmaları diğer algoritmalardan üstün sonuçlar vermişlerdir. Bu bulgular, NB algoritmalarının haber sınıflandırması için etkili ve güçlü bir yöntem olduğunu göstermektedir.

Özellikle Multinomial Naive Bayes ve varyasyonları, veri setinin doğasına uygun olarak en yüksek performansı sağlamıştır. Bunun nedeni, metin sınıflandırma problemlerinde kelime frekanslarını ve dağılımlarını etkili bir şekilde kullanabilme yeteneklerinde gizlidir.

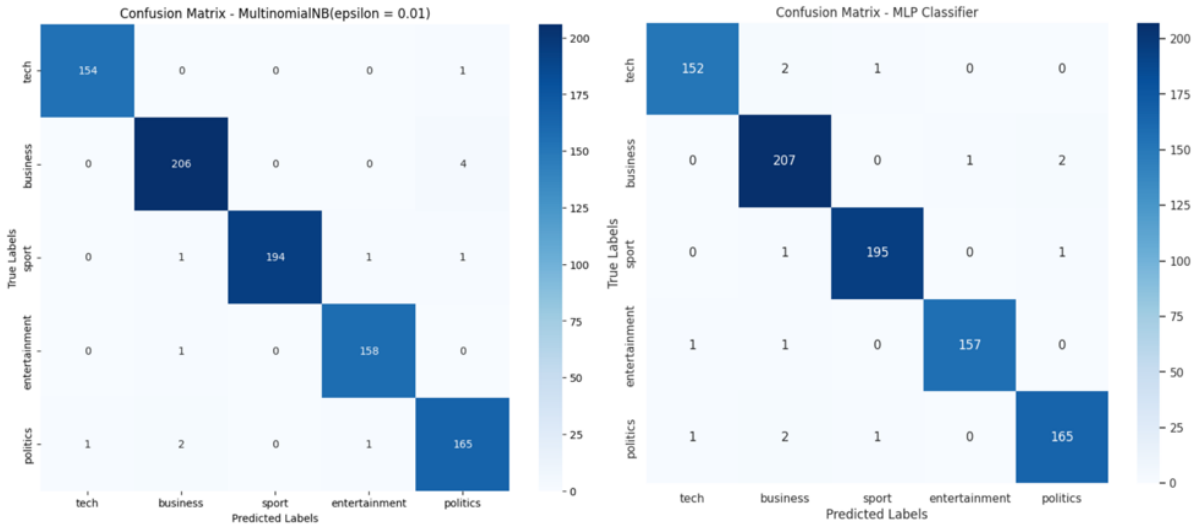
Epsilon değeri 0.01 olarak ayarlanmış Tree-Augmented Naive Bayes (TAN) algoritması, tüm metriklerde güçlü bir performans sergilemiştir. Doğruluk oranı %98.20, kesinlik %98.23, duyarlılık %98.20 ve F1 skoru %98.20. TAN'ın performansı, özellikler arasındaki bağımlılıkları modelleme

yeteneğinden kaynaklanmaktadır. Bu, özellik bağımsızlığını varsayan standart NB'in aksine, özellik etkileşimlerinin sınıflandırmada önemli bir rol oynadığı veri kümeleri için TAN'ı özellikle etkili kılar.



Şekil 5. MÖ Algoritmalarının Ortalama Performans Değerleri

Diğer makine öğrenmesi algoritmaları arasında ÇKA Sınıflandırıcısı, Lojistik Regresyon ve Doğrusal DVS en yüksek performansı sergilemişlerdir. Bu algoritmalar, sergiledikleri yakın performansları ile NB algoritmalarına alternatif olarak değerlendirilebilir. Ancak, Doğrusal Ayırtaç Analizi ve K-EYK gibi algoritmaların düşük performansları, bu algoritmaların haber sınıflandırma problemleri için uygun bir seçenek olmasını engellemektedir.



Şekil 6. Multinomial (epsilon = 0.01) ve ÇKA Sınıflandırıcı Algoritmalarının Karışıklık Matrisleri

Bu çalışmada ayrıca NB ve MÖ algoritmaları arasındaki en yüksek performans metrik değerlerine sahip ÇKA Sınıflandırıcısı ile Multinomial Naive Bayes (epsilon = 0.01) algoritmalarının performansları, oluşturulan *karışıklık matrisleri* (confusion matrices) yardımıyla karşılaştırma yoluna gidilmiştir. Karışıklık matrisi, bir modelin performansını değerlendirmek için kullanılan ve görsel olarak bir Çizelge şeklinde temsil edilen bir araçtır. Multinomial (epsilon = 0.01) ve ÇKA Sınıflandırıcı algoritmalarına ait karışıklık matrisleri Şekil 6'da sunulmaktadır. Her iki algoritma, teknoloji, iş, spor, eğlence ve politika kategorilerinde haber metinlerinin sınıflandırılması amacıyla kullanılmıştır.

Multinomial Naive Bayes algoritması için oluşturulan karışıklık matrisinde Teknoloji kategorisinde 154 doğru sınıflandırma yapılırken, 1 örnek politika kategorisinde olmak üzere yanlış sınıflandırılma yapılmıştır. İş kategorisinde ise 206 doğru sınıflandırma yapılırken, 1 örnek spor, 1

örnek eğlence ve 4 örnek politika kategorilerinde olmak üzere yanlış sınıflandırılma yapılmıştır. Benzer şekilde, spor kategorisinde 194 doğru sınıflandırma yapılırken, 1 örnek iş, 1 örnek eğlence ve 1 örnek politika kategorilerinde olmak üzere yanlış sınıflandırılma yapılmıştır. Devamında, eğlence kategorisinde 158 doğru sınıflandırma yapılırken, 1 örnek iş ve 1 örnek politika olmak üzere yanlış sınıflandırılma yapılmıştır. Son olarak politika kategorisi, 165 doğru sınıflandırma ile sonuçlanırken, 1 örnek teknoloji, 2 örnek iş ve 1 örnek eğlence kategorisinde olmak üzere yanlış sınıflandırılma ile sonuçlanmıştır.

ÇKA Sınıflandırıcı algoritması için oluşturulan karışıklık matrisinde Teknoloji kategorisinde 152 doğru sınıflandırma yapılmışken, 2 örnek iş, 1 örnek eğlence ve 1 örnek politika olmak üzere yanlış sınıflandırılma yapılmıştır. İş kategorisinde 207 doğru sınıflandırma yapılmışken, 1 örnek eğlence ve 2 örnek politika olmak üzere yanlış sınıflandırılmıştır. Spor kategorisinde ise 195 doğru sınıflandırma yapılırken, 1 örnek teknoloji ve 1 örnek politika olmak üzere yanlış sınıflandırılmıştır. Eğlence kategorisinde 157 doğru sınıflandırma yapılmışken, 1 örnek teknoloji ve 1 örnek iş olmak üzere yanlış sınıflandırılmıştır. Son olarak politika kategorisinde 165 doğru sınıflandırma ile karşılaşılmış iken, 1 örnek teknoloji, 2 örnek iş ve 1 örnek spor kategorisinde olmak üzere yanlış sınıflandırılma ile karşılaşılmıştır.

Her iki algoritmanın performansını karşılaştırırken, genel doğruluk oranlarının yanı sıra her bir kategorideki doğru ve yanlış sınıflandırmalar da incelenmiştir. ÇKA Sınıflandırıcısı, özellikle teknoloji ve iş kategorilerinde daha az hata yaparken, Multinomial Naive Bayes algoritması eğlence ve politika kategorilerinde nispeten daha iyi performans sergilemiştir. Her iki algoritmanın da belirli kategorilerdeki performansları arasında küçük farklar bulunmakla birlikte, genel olarak her ikisi de yüksek doğruluk oranlarına sahip etkili sınıflandırıcılar olarak değerlendirilmiştir.

Sonuç olarak, haber sınıflandırmasında NB algoritmalarının güçlü bir tercih olduğu, ancak belirli durumlarda diğer makine öğrenmesi algoritmalarının da etkili olabileceği rahatlıkla ifade edilebilir. Gelecekte yapılacak çalışmalarda, daha fazla veri ve farklı özellik mühendisliği yöntemleri kullanılarak bu algoritmaların performanslarının daha da artırılması mümkün olabilecektir.

SONUÇ

Bu çalışmada, BBC News Corpus'tan elde edilen haber başlıklarının oluşturduğu bir veri kümesinin kullanılmasıyla yetkin bir haber sınıflandırma sistemi oluşturmayı amaçlamış bulunuyoruz. Bu amaca hizmet üzere birçok NB varyantının yanı sıra diğer yaygın ve oldukça iyi bilinen MÖ algoritmalarının performansları, titiz ve detaylı bir inceleme ve araştırma aşamalarının ardından karşılaştırmalı olarak bir değerlendirmeye tabi tutulmuştur. Kapsamlı deneyler ve analizler sayesinde, bu algoritmaların metin sınıflandırma görevlerindeki etkinliğine ilişkin değerli bilgiler edinilmiştir. Performans değerlendirmesi için doğruluk, kesinlik, duyarlılık ve F1 skoru metrikleri kullanılmıştır. Elde edilen sonuçlar, her iki algoritma grubu arasında belirgin performans farklılıkları olduğunu ortaya koymuştur.

NB algoritmalarının, haber sınıflandırması problemlerinde genel olarak yüksek bir performans sergilediği gözlemlenmiştir. Gaussian, Multinomial, Complement, Bernoulli ve TAN gibi farklı NB türleri incelendiğinde, özellikle Multinomial Naive Bayes algoritmasının üstün performansı dikkat çekmektedir. Multinomial Naive Bayes ve varyasyonları, doğruluk, kesinlik, duyarlılık ve F1 skoru metriklerinde %98'in üzerinde sonuçlar üretmiştir. Bu başarı, Multinomial Naive Bayes algoritmasının metin sınıflandırma problemlerinde kelime frekanslarını etkili bir şekilde kullanabilme yeteneğine bağlanabilir.

Özellikle Multinomial Naive Bayes (Alpha, 0.5) varyasyonu, %98.53 doğruluk, %98.53 kesinlik, %98.53 duyarlılık ve %98.53 F1 skoru ile en yüksek performansı sergilemiştir. Bu sonuçlar, NB algoritmalarının özellikle büyük veri kümelerinde ve yüksek boyutlu özellik alanlarında etkili olabileceğini göstermektedir.

Diğer Naive Bayes varyantlarıyla karşılaştırıldığında, TAN'ın doğruluğu, en iyi performans gösteren Multinomial Naive Bayes yapılandırmalarıyla (örneğin, Alpha 0.5 ile Multinomial ve epsilon 0.01 ile Multinomial) paraleldir ve bu yapılandırmalar da yaklaşık %98.53 doğruluk göstermektedir. Ancak, TAN'ın kesinliği (%98.23), Multinomial (epsilon = 0.01) tarafından elde edilen en yüksek kesinliğin (%98.55) biraz gerisinde kalmaktadır. TAN'ın duyarlılık ve F1 skoru, kesinlik ile yakından uyum içindedir ve bu, kesinlik ve duyarlılık arasında önemli bir denge kaybı olmadan dengeli bir performans sergilediğini göstermektedir. Bu denge, yanlış pozitif ve yanlış negatiflerin benzer maliyet taşıdığı uygulamalar için kritiktir.

Diğer makine öğrenmesi algoritmalarının performansı incelendiğinde, genellikle NB algoritmalarından daha düşük sonuçlar elde edilmiştir. Lojistik Regresyon, Rastgele Orman, Doğrusal DVS ve ÇKA Sınıflandırıcısı gibi algoritmalar yüksek performans sergilemişler ve Naive Bayes algoritmalarına yakın sonuçlar vermişlerdir. Özellikle ÇKA Sınıflandırıcısı, %98.31 doğruluk, %98.32 kesinlik, %98.31 duyarlılık ve %98.31 F1 skoru ile en yüksek performansı sergilemiş bulunmaktadır.

Bununla birlikte, Doğrusal Ayırtaç Analizi ve K-EYK gibi algoritmaların performansları oldukça düşük kalmıştır. Bilhassa Doğrusal Ayırtaç Analizi algoritması, %38.20 doğruluk yüzdesi ile en düşük performansı sergileyen algoritma olmuştur.

NB algoritmalarının, özellikle metin sınıflandırma problemlerinde üstün performans göstermeleri, bu algoritmaların haber sınıflandırması gibi DDİ görevlerinde güçlü bir tercih olduğunu ortaya koymaktadır. NB algoritmaları, basit ve hızlı olmaları nedeniyle büyük veri kümelerinde ve gerçek-zaman uygulamalarında kullanılabilir.

Diğer MÖ algoritmaları arasında Lojistik Regresyon, Rastgele Orman, Doğrusal DVS ve ÇKA Sınıflandırıcısı gibi algoritmalar, NB algoritmalarına yakın performansları ile pekala alternatif algoritmalar olarak değerlendirilebilir. Bu algoritmalar ile özellikle doğruluk, kesinlik ve duyarlılık ve F1 skoru metriklerinde yüksek sonuçlar elde edilmiştir. Ancak, Doğrusal Ayırtaç Analizi ve K-EYK gibi algoritmaların düşük performansları, bu tür algoritmaların haber sınıflandırması problemlerinde kullanılmasını engellemektedir.

Bu çalışmanın sonuçları, haber sınıflandırması problemlerinde Naive Bayes algoritmalarının güçlü bir tercih olduğunu bize göstermektedir. Ancak, gelecekteki çalışmalarda, daha fazla veri ve farklı özellik mühendisliği yöntemleri kullanılarak bu algoritmaların performanslarını daha da artırmak mümkün olabilecektir. Özellikle derin öğrenme yöntemlerinin ve karma modellerin kullanılması, haber sınıflandırması uygulamalarında performans artırıcı bir unsur olabilir. Yakın gelecekte, bu çalışmada kullanılan veri kümesi ve özelliklerin, algoritmaların performansını nasıl etkilediği daha ayrıntılı olarak incelenecektir. Farklı dil işleme teknikleri, kelime temsil yöntemleri ve özellik seçimi yöntemleri, sınıflandırma performansını önemli ölçüde etkileyebilir. Bu nedenle, gelecekteki çalışmalarda, bu faktörlerin etkisini incelemek ve bilahare optimize etmek suretiyle sınıflandırma performansını daha da artırmak mümkün olabilecektir.

Sonuç olarak, bu çalışma, farklı NB ve diğer bilinen makine öğrenmesi algoritmalarının haber sınıflandırma sistemleri üzerindeki performanslarını karşılaştırmış ve NB algoritmalarının bu tür problemler için güçlü bir tercih olduğunu ortaya koymuştur. Ancak, farklı algoritmaların ve tekniklerin

kullanımı, belirli durumlarda daha iyi sonuçlar verebilir ve bu nedenle her bir problem için uygun algoritma ve yöntemlerin dikkatli bir şekilde seçilmesi önem arz etmektedir.

Çıkar Çatışması

Makale yazarları aralarında herhangi bir çıkar çatışması olmadığını beyan ederler

Yazar Katkısı

Yazarlar makaleye eşit oranda katkı sağlamış olduklarını beyan eder.

KAYNAKLAR

Albahr, A., & Albahar, M. (2020). An empirical comparison of fake news detection using different machine learning algorithms. International Journal of Advanced Computer Science and Applications, 11(9).

Bracewell, D. B., Yan, J., Ren, F., & Kuroiwa, S. (2009). Category classification and topic discovery of japanese and english news articles. Electronic Notes in Theoretical Computer Science, 225, 51-65..

Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. Knowledge-Based Systems, 192, 105361.D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

Granik, M., & Mesyura, V. (2017, May). Fake news detection using naive Bayes classifier. In 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON) (pp. 900-903). IEEE.

Greene, D., & Cunningham, P. "BBC Datasets," 2006. [Online]. Available: http://mlg. ucd.ie/datasets/ bbc.html.

Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), 381-386.M. I. Rana, S. Khalid, and M. U. Akbar, "News classification based on their headlines: A review," in 17th IEEE International Multi Topic Conference 2014, Karachi, Pakistan, 2014: IEEE, pp. 211-216.

Patel, A., & Meehan, K. (2021, June). Fake news detection on reddit utilising countvectorizer and term frequency-inverse document frequency with logistic regression, multinominalnb and support vector machine. In 2021 32nd Irish signals and systems conference (ISSC) (pp. 1-6). IEEE.M. M. Saritas and A. Yasar, "Performance analysis of ANN and Naive Bayes classification algorithm for data classification," *International journal of intelligent systems and applications in engineering*, vol. 7, no. 2, pp. 88-91, 2019.

Shahi, T. B., & Pant, A. K. (2018, February). Nepali news classification using Naive Bayes, support vector machines and neural networks. In 2018 international conference on communication information and computing technology (iccict) (pp. 1-5). IEEE.A. N. Chy, M. H. Seddiqui, and S. Das, "Bangla news classification using naive Bayes classifier," in 16th Int'l Conf. Computer and Information Technology, Khulna, Bangladesh, 2014: IEEE, pp. 366-371.

Sristy, N. B., & Somayajulu, D. V. L. N. (2012, December). Semi-supervised Learning of Naive Bayes Classifier with feature constraints. In Proceedings of the First International Workshop on Optimization Techniques for Human Language Technology (pp. 65-78).