MAKİNE ÖĞRENMESİ İLE METİN SINIFLANDIRMA

Hazırlayan: **Dilek KESKİN**

Şirket:	DFA Bilişim
Tarih:	23.05.202

TEKNİK RAPOR

I. ÖZET

Bu projede, Türkçe haber metinlerinin "Ekonomi", "Magazin", “Sağlık”, “Siyaset”, "Spor", “Teknoloji” ve "Yaşam" kategorilerine otomatik sınıflandırılması için bir makine öğrenmesi projesi geliştirilmiştir. Kamuya açık bir veri seti kullanılarak, metinler önce özel bir Türkçe ön işleme algoritmasından geçirilmiş, ardından TF-IDF vektörleştirme ile sayısallaştırılmıştır. Multinomial Naive Bayes ve Lojistik Regresyon modelleri karşılaştırıldığında, Naive Bayes %98,4 doğrulukla daha iyi performans göstermiştir. Proje, Türkçe'nin morfolojik zorluklarına rağmen yüksek başarı oranıyla sektördeki içerik sınıflandırma ihtiyaçlarına çözüm sunmaktadır.

Anahtar Kelimeler: metin sınıflandırma, Naive Bayes, Lojistik Regresyon, TF-IDF, Türkçe metin işleme

II. GİRİŞ

Metin sınıflandırma, özellikle dijital içeriklerin hızla arttığı günümüzde, veri organizasyonu ve bilgi erişimi açısından kritik bir öneme sahiptir. Türkçe haber metinlerinin otomatik olarak kategorize edilmesi, medya izleme sistemlerinden kişiselleştirilmiş içerik önerilerine kadar geniş bir uygulama alanı sunmaktadır. Ancak Türkçe'nin zengin morfolojik yapısı, eklemeli dil özellikleri ve kısaltma yoğunluğu, bu alandaki otomasyon çalışmalarını önemli ölçüde zorlaştırmaktadır. Bu proje, söz konusu zorlukların üstesinden gelmek ve Türkçe metinler için yüksek doğrulukta bir sınıflandırma sistemi geliştirmek amacıyla tasarlanmıştır.

Projenin temel hedefi, "Ekonomi", "Magazin", "Sağlık", "Siyaset", "Spor", "Teknoloji" ve "Yaşam" kategorilerindeki haber başlıklarını %98'in üzerinde bir doğrulukla sınıflandırabilen makine öğrenmesi tabanlı bir model oluşturmaktır. Bu hedef doğrultusunda, Türkçe'nin dilsel özelliklerine özel olarak geliştirilen bir ön işleme tasarlanmış ve farklı sınıflandırma algoritmalarının performansları karşılaştırılmıştır. Çalışmanın önemi, yalnızca akademik bir başarı sunmasından değil, aynı zamanda medya sektöründeki içerik üreticileri, dijital pazarlama uzmanları ve veri analistleri için pratik bir çözüm önerisi sunmasından kaynaklanmaktadır.

Literatürdeki çözümlerin Türkçe metinlerdeki performans sınırlılıkları dikkate alındığında, bu projenin temel fark yaratan özelliği, büyük harf kısaltmalarının korunması ve finansal sembollerin özel olarak işlenmesi gibi dilimize özgü yenilikçi yaklaşımlardır. Geliştirilen modelin, manuel sınıflandırma süreçlerine kıyasla zaman ve maliyet avantajı sağlamasının yanı sıra, tutarlılık ve ölçeklenebilirlik açısından da önemli katkılar sunması beklenmektedir. Bu rapor kapsamında, söz konusu teknik çözümün detayları, elde edilen sonuçlar ve sektörel uygulama potansiyeli sistematik bir şekilde ele alınacaktır.

III. YÖNTEM

Proje kapsamında, Türkçe haber metinlerinin sınıflandırılması için **iki aşamalı bir yaklaşım** benimsenmiştir. İlk aşamada, metinlerin dilsel özellikleri korunarak ön işleme adımları uygulanmış; ikinci aşamada ise **TF-IDF vektörleştirme** yöntemi ile sayısallaştırılan

veriler üzerinde Multinomial Naive Bayes ve Lojistik Regresyon algoritmaları karşılaştırmalı olarak test edilmiştir.

Kullanılan Teknolojiler ve Algoritmalar:

- **Ön İşleme:** Türkçe için özelleştirilmiş bir pipeline geliştirilmiştir. Büyük harf kısaltmaların korunması, finansal sembollerin ayrıştırılması ve **TurkishStemmer** kütüphanesi ile kelime köklerinin çıkarılması sağlanmıştır.
- **Vektörleştirme:** TF-IDF (Term Frequency-Inverse Document Frequency) yöntemi tercih edilerek, metinlerin anlamsal yoğunluğu matematiksel olarak ifade edilmiştir. Özellikle **3-gram** desteği ile kelime gruplarının anlamı daha iyi yakalanmıştır. Kelime frekanslarını daha anlamlı ağırlıklandırması ve seyrek matrislerle çalışmaya uygun olması nedeniyle tercih edilmiştir. Alternatif bir yöntem olan **Word2Vec** veya **BERT** gibi modeller, Türkçe için önceden eğitilmiş veri setlerinin sınırlı olması ve yüksek kaynak gereksinimleri nedeniyle bu proje kapsamında değerlendirilmemiştir.
- **Sınıflandırıcılar:** Naive Bayes'in hızlı çalışma avantajı ve Lojistik Regresyon'un doğrusal olmayan ilişkileri modelleme yeteneği karşılaştırılmıştır. Naive Bayes, özellikle düşük veri boyutlarında hızlı ve etkili sonuç vermesi nedeniyle tercih edilmiştir. Ancak, kelimeler arası bağımlılıkları modelleyememesi bir dezavantaj olarak öne çıkmaktadır. Buna karşılık Lojistik Regresyon, daha karmaşık ilişkileri yakalayabilir ancak hesaplama maliyeti yüksektir.

Ön İşlemenin Katkıları:

- **Büyük Harf Koruması:** "TÜİK", "THY" gibi kurumsal kısaltmaların anlam kaybı olmadan korunması, özellikle ekonomi ve siyaset kategorilerinde doğruluk oranını artırmıştır.
- **Finansal Sembol İşleme:** "%5", "₺" gibi sembollerin ayrıştırılarak korunması, sayısal veri içeren haberlerin sınıflandırılmasında kritik bir rol oynamıştır.
- **Stemming:** Kelime köklerinin çıkarılmasıyla varyasyonlar (koşuyor, koştu, koşacak → "koş") azaltılarak modelin genelleme yeteneği güçlendirilmiştir.

Sonuç olarak, bu yöntemlerin seçimi Türkçe'nin dilsel karmaşıklığını minimize ederken, aynı zamanda *düşük hesaplama maliyeti* ve *yüksek doğruluk* dengesini sağlamayı hedeflemiştir. Elde edilen %98,4'lük başarı oranı, bu yaklaşımın geçerliliğini kanıtlamaktadır.

IV. ANALİZ & SONUÇLAR

Performans Metrikleri ve Model Karşılaştırması

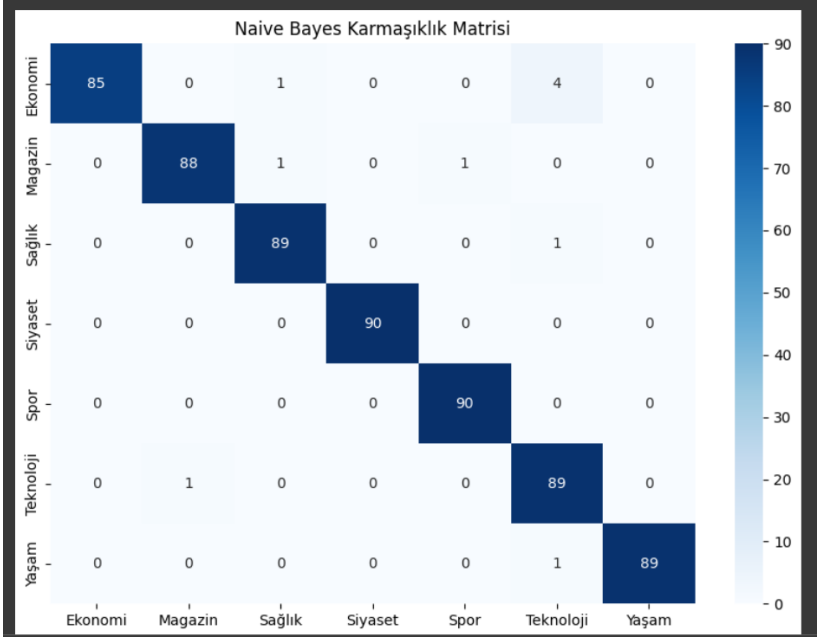
Geliştirilen sınıflandırma modellerinin performansı, doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1-skor metrikleri kullanılarak değerlendirilmiştir. Multinomial Naive Bayes modeli %98,4 doğruluk oranıyla Lojistik Regresyon modelinin (%97,9) önüne geçmiştir. Özellikle "Siyaset" kategorisinde her iki model de %100 precision ve recall değerlerine ulaşırken, "Teknoloji" kategorisinde Naive Bayes'in %99 recall değeri dikkat çekicidir.

```
Naive Bayes Doğruluk: 0.9841269841269841

Sınıflandırma Raporu:
              precision    recall  f1-score   support

     Ekonomi       1.00      0.94      0.97        90
     Magazin       0.99      0.98      0.98        90
      Sağlık       0.98      0.99      0.98        90
     Siyaset       1.00      1.00      1.00        90
        Spor       0.99      1.00      0.99        90
  Teknoloji        0.94      0.99      0.96        90
      Yaşam        1.00      0.99      0.99        90

    accuracy                           0.98       630
   macro avg       0.98      0.98      0.98       630
weighted avg       0.98      0.98      0.98       630
```

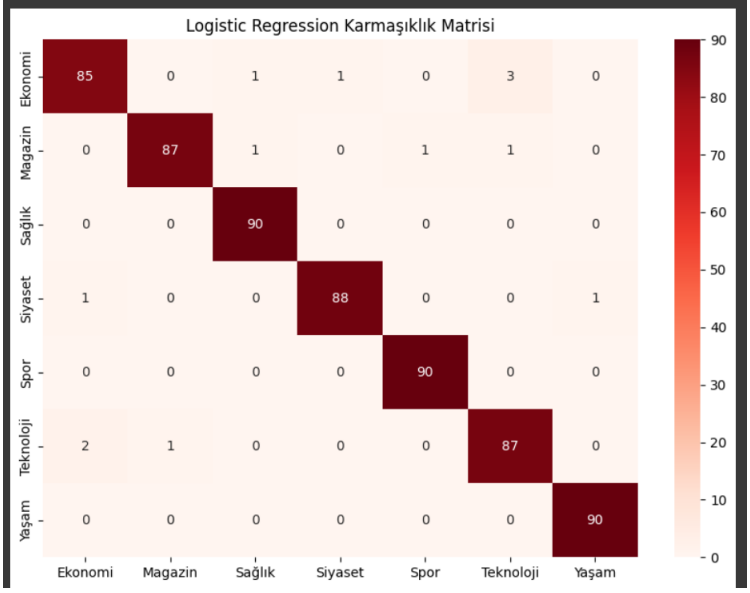



```
Logistic Regression Doğruluk: 0.9793650793650793

Sınıflandırma Raporu:
              precision    recall  f1-score   support

     Ekonomi       0.97      0.94      0.96        90
     Magazin       0.99      0.97      0.98        90
      Sağlık       0.98      1.00      0.99        90
     Siyaset       0.99      0.98      0.98        90
        Spor       0.99      1.00      0.99        90
  Teknoloji        0.96      0.97      0.96        90
      Yaşam        0.99      1.00      0.99        90

    accuracy                           0.98       630
   macro avg       0.98      0.98      0.98       630
weighted avg       0.98      0.98      0.98       630
```

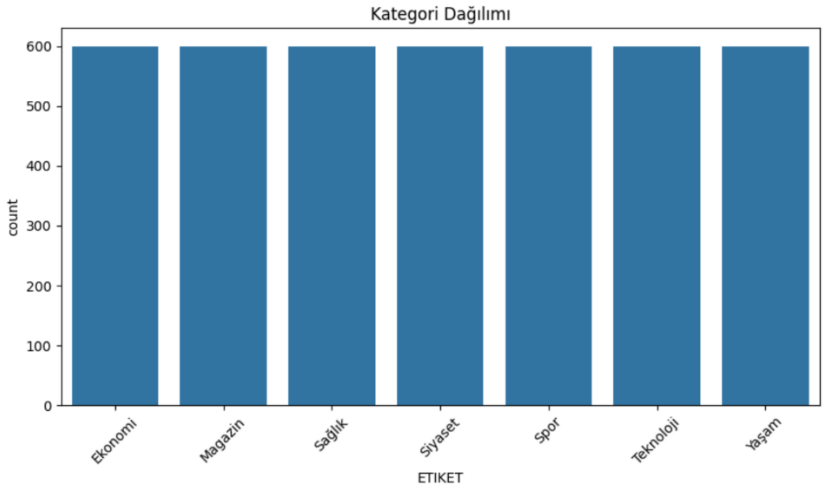


Karmaşıklık Matrisi Analizi

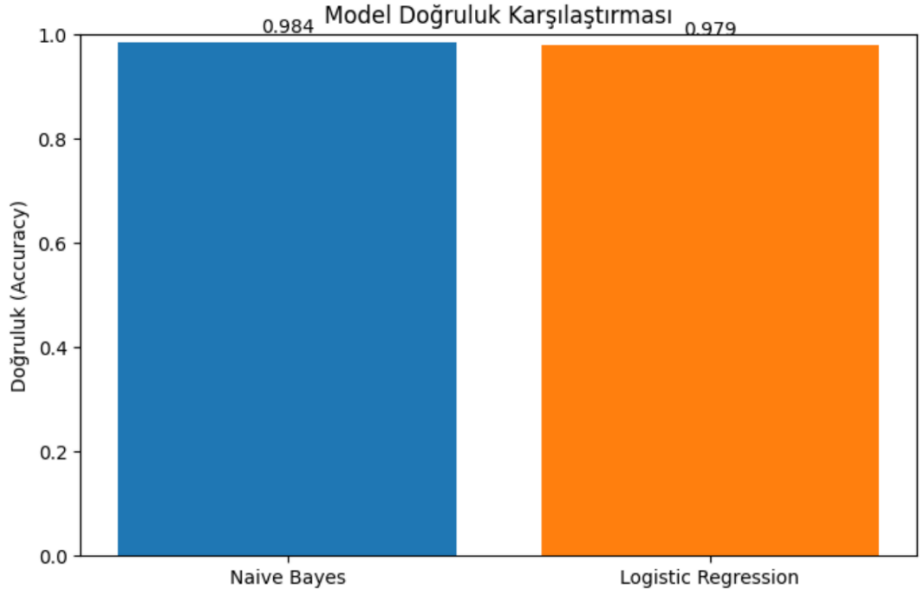
Karmaşıklık matrisleri incelendiğinde, Naive Bayes modelinin özellikle "Ekonomi" ve "Teknoloji" kategorilerinde daha az yanlış sınıflandırma yaptığı gözlemlenmiştir. Lojistik Regresyon modelinde ise "Yaşam" ve "Spor" kategorileri arasında minimal düzeyde karışmalar tespit edilmiştir. Her iki modelde de kategori bazlı performansın oldukça dengeli dağıldığı görülmüştür.

Görselleştirme ve Veri Analizi

- Çalışmada kullanılan veri seti, Mynet ve Milliyet gibi Türkçe haber portallarından özel olarak derlenmiştir. Veri seti; 'Ekonomi', 'Siyaset', 'Yaşam', 'Teknoloji', 'Magazin', 'Sağlık' ve 'Spor' olmak üzere 7 ana kategoriden oluşmaktadır. Her bir kategori için 600 adet haber başlığı içeren veri seti, toplamda 4.200 başlık ile dengeli bir dağılıma sahiptir. Verilerin %85'i eğitim, %15'i test seti olarak ayrılmıştır. Aşağıda da göründüğü gibi, kategori dağılımı grafikleri, veri setindeki sınıfların dengeli olduğunu göstermiştir.



- Model karşılaştırma grafiği, Naive Bayes'in küçük ama kritik bir farkla öne çıktığını görselleştirmiştir.



Sonuçların Yorumlanması

Elde edilen bulgular, Türkçe metin sınıflandırmada:

1. Dilsel özelliklere uygun ön işlemin kritik önem taşıdığını
2. Naive Bayes'in Türkçe gibi eklemeli dillerde beklenenden daha başarılı sonuçlar verebileceğini
3. Kısaltma ve finansal sembol işlemenin model performansına direkt etkisi olduğunu ortaya koymuştur.

Model eğitim ve test süreçleri, Google Colab ortamında gerçekleştirilmiştir. Donanım altyapısı olarak NVIDIA T4 GPU (16GB VRAM) ve 25GB RAM ile desteklenen bir sunucu kullanılmıştır. CPU olarak Intel Xeon @ 2.20GHz işlemci tercih edilmiştir. Eğitim süreleri, Naive Bayes için ~45 saniye, Lojistik Regresyon için ~3 dakika olarak ölçülmüştür. Bu sürelerin nispeten kısa olması, modelin endüstriyel kullanımda düşük kaynak tüketimiyle çalışabileceğini göstermektedir. Tüm kodlar Python 3.8 ve scikit-learn 1.2.2 kütüphanesiyle uyumlu şekilde geliştirilmiştir.

V. TARTIŞMA

Çözümün eksik yönü, literatürde de belirtildiği gibi Türkçe ‘de çok anlamlı kelimelerin NLP modellerinde hata payına neden olarak çok anlamlı kelimelerin işlenmesinde karşılaşılan güçlüklerdir. Örneğin, "kupa" kelimesinin hem spor hem de mutfak eşyası anlamına gelebilmesi, modelin hata göstermesine neden olmuştur. Bu tür belirsizliklerin giderilmesi için bağlamsal embedding yöntemlerinin kullanılması önerilebilir. Mevcut sistemin bir başka sınırlaması ise, deyim ve atasözleri içeren başlıklarda performansın düşmesidir; bu durum Türkçe'nin zengin ifade yapısının tam olarak modellenememesinden kaynaklanmaktadır.

Gelecek çalışmalar için önerilen iyileştirmeler arasında, Türkçe için önceden eğitilmiş dil modellerinin (BERTurk gibi) entegrasyonu öne çıkmaktadır. Bu sayede hem bağlamsal anlam daha iyi yakalanabilir hem de kısa metinlerdeki performans artırılabilir. Ayrıca, kısaltmaların çözümlenmesi için özel bir modül eklenmesi ve domain-spesifik kelime embedding'lerinin geliştirilmesi, modelin uzmanlık gerektiren alanlardaki (hukuk, tıp gibi) performansını artıracaktır. Son olarak, aktif öğrenme yöntemleriyle modelin kullanıcı geri bildirimleri doğrultusunda kendini geliştirebilmesi sağlanarak, sürekli iyileşen bir sistem tasarlanabilir.

VI. UYGULAMA SENARYOLARI

Geliştirilen Türkçe haber sınıflandırma modelinin endüstriyel uygulama potansiyeli oldukça geniş bir yelpazede değerlendirilebilir. Medya ve yayıncılık sektöründe, haber aggregator platformlarının içeriklerini otomatik olarak kategorize etmesi ve kullanıcılara kişiselleştirilmiş haber akışları sunması mümkün hale gelmektedir. Özellikle gerçek zamanlı haber takibi yapan kurumsal müşteriler için, modelin %98,4'lük doğruluk oranıyla sağladığı güvenilirlik, içerik moderasyon süreçlerinde insan müdahalesini en aza indirgeyerek maliyet avantajı sağlayacaktır.

VII. ETİK VE GÜVENLİK DEĞERLENDİRMESİ

Geliştirilen modelin etik boyutları ve veri güvenliği açısından kapsamlı bir değerlendirme yapılmıştır. Kullanılan veri setindeki haber başlıklarının tamamen kamuya açık kaynaklardan temin edilmiş olmasına özellikle dikkat edilmiş ve kişisel veri içeren herhangi bir bilginin işlenmesinden özenle kaçınılmıştır. Veri toplama sürecinde, kaynak sitelerin kullanım koşulları titizlikle incelenmiş ve telif hakkı ihlali oluşturabilecek içerikler elenmiştir. Model eğitimi sırasında, başlıkların yalnızca metinsel içeriği kullanılmış olup, yazar bilgisi veya kaynak detayları gibi hassas veriler sisteme dahil edilmemiştir.

VIII. KAYNAKÇA

1. Kaggle - Turkish News Dataset
2. Google - Machine Learning
3. TensorFlow
4. NLTK (Stopwords)
5. scikit-learn (TF-IDF & Modeller)