**MACHİNE LEARNİNG ALGORİTMALARI İLE TWEET’LERİN OLUMLU VEYA OLUMSUZ OLDUĞUNU ANLAYABİLME ÜZERİNE BİR ÇALIŞMA**

Kaan Berk Duman, Ertuğrul Sungur, Esra Güneş, İlayda Belen

**Kaan Berk Duman**

**İstanbul Medeniyet Üniversitesi Mühendislik ve Doğa Bilimleri Fakültesi Bilgisayar Mühendisliği Bölümü**

**Abstract**

Derin öğrenme, yapay zekâ ve makine öğrenmesi alanlarının önemli bir konusu haline gelmiştir. Özellikle son yıllarda, farklı derin öğrenme yöntemleri öneren çalışmaların sayısı hızla artmıştır. Doğal dil işlemenin çeşitli alt alanlarında da bu yöntemler yaygın olarak kullanılmaktadır. Bu çalışmada derin öğrenme yaklaşımları kullanılarak sosyal medyada atılan tweet’lerin olumlu veya olumsuz bir yorum olup olmadığını anlaması için bir yapay zekâ geliştirilmeye çalışılmıştır.

**Anahtar kelimeler:** Derin Öğrenme, Doğal Dil İşleme, Makine Öğrenmesi.

1. **Introduction**

Derin öğrenme, makine öğrenmesinin bir alt dalıdır ve özellikle yapay sinir ağlarına dayanan bir yöntemdir. Derin öğrenme, büyük veri kümeleri üzerinde çalışarak yüksek seviyede özellik çıkarımı ve otomatik öğrenme sağlayan çok katmanlı sinir ağlarını kullanır. Bu yöntem, verinin doğrudan işlenmesine ve çok karmaşık soyut ilişkilerin modellenmesine olanak tanır. Derin öğrenme, geleneksel makine öğrenmesi tekniklerinden farklı olarak, özellik çıkarım işlemini otomatikleştirir ve veriden doğrudan öğrenir. [1]

Derin öğrenmede kullanılan temel yöntemler ve teknikler şunlardır:[2]

1. **Yapay Sinir Ağları:** Temel derin öğrenme yapı taşlarıdır ve genellikle çok katmanlıdır.
2. **Derin Sinir Ağları:** Birden fazla gizli katmana sahip olan sinir ağlarıdır. Her katman, bir önceki katmandan gelen bilgiyi işler ve daha soyut özellikler çıkarır.
3. **Evrişimsel Sinir Ağları:** Özellikle görüntü işleme alanında yaygın olarak kullanılır. CNN’ler, uzaysal hiyerarşiyi öğrenmek için evrişim ve havuzlama katmanlarını kullanır.
4. **Tekrarlayan Sinir Ağları:** Zaman serisi verilerini ve ardışık verileri işlemek için kullanılır. RNN’ler, önceki durum bilgilerini kullanarak sıralı verilerdeki bağımlılıkları öğrenir.
5. **Uzun Kısa Süreli Bellek:** RNN’lerin bir çeşididir ve uzun vadeli bağımlılıkları daha iyi öğrenmek için özel olarak tasarlanmıştır.
6. **Otoenkoderler:** Verinin sıkıştırılması ve gürültüsünün azaltılması için kullanılır. Otoenkoderler, giriş verisini sıkıştırır ve daha sonra yeniden oluşturur.

Derin öğrenme, geniş bir uygulama yelpazesinde kullanılmaktadır:[3]

1. **Görüntü İşleme:** Görüntü sınıflandırma, nesne tespiti, yüz tanıma, tıbbi görüntü analizi.
2. **Doğal Dil İşleme:** Makine çevirisi, metin sınıflandırma, konuşma tanıma, duygu analizi.
3. **Ses İşleme:** Konuşma tanıma, ses sentezi, müzik öneri sistemleri.
4. **Otonom Araçlar:** Görüntü işleme, yol ve engel tespiti, karar verme sistemleri.
5. **Sağlık:** Hastalık teşhisi, genetik veri analizi, kişisel tedavi planları.
6. **Finans:** Dolandırıcılık tespiti, hisse senedi fiyat tahmini, müşteri hizmetleri otomasyonu.
7. **Eğlence:** Film ve müzik öneri sistemleri, oyun geliştirme.

Derin öğrenme, büyük veri kümeleri ve güçlü hesaplama kaynakları sayesinde birçok alanda devrim niteliğinde gelişmeler sağlamaktadır. Bu yöntemlerin uygulanmasıyla, çeşitli sektörlerde verimlilik artışı ve yeni hizmetler sağlanabilmektedir. [4]

Yukarıdaki listede ikinci madde olarak yer alan doğal dil işleme konusu, yapay zekâ alanı içerisinde değerlendirilmektedir ve metin madenciliği olarak da adlandırılmaktadır. Doğal dil işleme, doğal dilde oluşturulmuş metinler üzerinde otomatik olarak gerçekleştirilen metin ayrıştırma, metin sınıflandırma gibi birçok önemli konuyu içinde barındırmaktadır ve uygulama alanları gün geçtikçe artmaktadır. [5]

Bu çalışmamızda; doğal dil işleme problemlerinin çözümü için makine öğrenmesi yöntemlerini kullanarak olumlu ve olumsuz tweet’leri ayırmaya çalıştık ve araştırmacıların dikkatine sunduk. Bu makalenin katkıları şu şekilde sıralanabilir:

* İleride çalışmamızın konusu olan tweet’lerin olumlu veya olumsuz olduğunu algılama alanında ilerlemek isteyen için bir ön kaynaktır.
* Yapay zekâ veya doğal dil işlemeye yeni başlayanlar için örnek bir çalışma olacaktır.

1. **Related Works**

Bir önceki bölümde anlatılan farklı yöntemlerden özellikle yinelenen sinir ağları ve evrişimsel sinir ağlarının doğal dil işleme problemlerini çözmede de başarılı sonuçlar verdiği görülmüştür. Bu problemler genel olarak duygu analizi çıkarma, olay çıkarımı, sözcük türü etiketleme, metin sıralama, otomatik harf çevirisi ve otomatik soru cevaplama gibi problemlerdir. [6]

1. **Duygu Analizi:** Duygu analizi, verilen bir metindeki duygu ve fikrin otomatik olarak belirlenmesi olarak tanımlanmaktadır.
2. **Olay Çıkarımı:** Doğal dilde oluşturulmuş metinlerde bahsedilen olayların otomatik çıkarımı da bilgi çıkarımı konusunun bir alt alanıdır ve dolayısıyla doğal dil işlemenin bir problemidir.
3. **Sözcük Türü Etiketleme:** Sözcük türü etiketleme, doğal dilde verilen bir metindeki kelimelerin türlerinin sınıflandırılması problemidir.
4. **Metin Sıralama:** Metin sıralama, bilgi erişiminin kapsamında bulunan önemli bir araştırma konusudur. Bu sıralama işlemi için de derin öğrenme yöntemini kullanan çalışmalar mevcuttur.
5. **Otomatik Harf Çevirisi:** Otomatik çeviri alanında kullanılan otomatik harf çevirisi ile kelimelerin bir alfabedeki harflerden diğer alfabedeki harflere çevrilmesini sağlamaktadır.
6. **Otomatik Soru Cevaplama**: Otomatik soru cevaplama sistemleri, doğal dilde ifade edilmiş bir sorunun cevabını dildeki büyük veri kümeleri içerisinde bularak yine doğal dilde otomatik olarak cevaplamayı amaçlayan sistemlerdir.
7. **Background/ Metodoloji**

## 3.1 Text Representation

Text representation, doğal dilde yazılmış metinlerin sayısal bir forma dönüştürülmesi işlemidir. Bu sayısal temsil, bilgisayarların metin verilerini işleyebilmesini ve analiz edebilmesini sağlar. [7]

### 3.1.1. word2vec

Word2Vec, doğal dil işleme alanında kullanılan bir kelime gömme yöntemidir. Google tarafından 2013 yılında geliştirilen Word2Vec, kelimeleri düşük boyutlu yoğun vektörlerle temsil ederek kelimeler arasındaki anlamsal ilişkileri yakalar.[8]

### 3.1.2 TF-IDF

Bir kelimenin belgede ne kadar önemli olduğunu belirtmek için kullanılır. TF, bir kelimenin bir belgede ne sıklıkla geçtiğini gösterir. IDF, bir kelimenin tüm belgeler arasında ne kadar nadir olduğunu ölçer. [9]

### 3.1.3 BOW

Metin belgesindeki kelimelerin sıklıklarına dayanan basit bir temsildir. Metin içindeki her kelimenin sayısını tutan bir vektör oluşturur. Kelimelerin sırası ve dilbilgisel yapısını dikkate almaz.[10]

## 3.2. Data set

|  |  |  |
| --- | --- | --- |
| TARGET | TWEET | SCORE |
| True | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | "@POTUS Biden Blunders – 6 Month Update | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | Inflation, Delta mismanagement, COVID for kids, Abandoning Americans in Afghanistan, Arming the Taliban, S. Border crisis, Breaking job growth, Abuse of power (Many Exec Orders, $3.5T through Reconciliation, Eviction Moratorium)...what did I miss?" | | | | | | | | | | | | | | | | | | | | | | | | 5 |
| True | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | "@S0SickRick @Stairmaster\_ @6d6f636869 Not as many people are literally starving and out in the streets as they were in the 19th century. Isnt capitalism grand? | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | Meanwhile, were facing an eviction moratorium threatening to make millions of Americans homeless. | | | | | | | | | |  |  |  |  |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | Fuck off with this corporatist propaganda." | | | | |  |  |  |  |  |  |  |  |  |  | | 3 |
| True | THE SUPREME COURT is siding with super rich property owners and over poor struggling AMERICANS by blocking the eviction moratorium during a pandemic which isn't even over yet. ASSHOLES! get ready for more homeless people! | 4 |
| ……… | ………….. | …….. |

Tablo 1 - Data Set

* 1. **Machine learning**

Makine öğrenmesi, bilgisayarların açıkça programlanmadan verilerden öğrenmesini sağlayan bir yapay zekâ alt disiplinidir. Makine öğrenmesi algoritmaları, örnek verileri kullanarak belirli görevlerde performanslarını arttırmak için modeller oluşturur ve bu modeller yeni verilere uygulandığında tahminler veya kararlar üretir.[11]

Bu çalışmada, makine öğrenimi yöntemleri arasından KNN, Decision Tree ve NB kullanılmıştır. Bu sınıflandırıcılar bu bölümde açıklanmaktadır.

* + 1. **Naïve Bayes**

Naive Bayes, olasılık teorisine dayanan basit ama güçlü bir makine öğrenmesi algoritmasıdır. Genellikle sınıflandırma görevlerinde kullanılır. Naive Bayes sınıflandırıcısı, her özelliğin bağımsız olduğunu varsayar ve bu bağımsızlık varsayımı altında sınıf olasılıklarını hesaplar.[12]

* + 1. **K-nearest Neighbor**

K-En Yakın Komşu (KNN) algoritması, denetimli öğrenme yöntemlerinden biri olan ve hem sınıflandırma hem de regresyon problemlerinde kullanılan basit ve sezgisel bir makine öğrenmesi algoritmasıdır. KNN, bir veri noktasının sınıfını veya değerini belirlemek için bu veri noktasına en yakın k komşunun sınıflarına veya değerlerine bakar.[13]

* + 1. **Decision Tree**

Karar Ağaçları hem sınıflandırma hem de regresyon problemlerinde kullanılan popüler ve sezgisel bir makine öğrenmesi algoritmasıdır. Karar ağaçları, karar verme sürecini ağaç yapısına benzer şekilde modelleyerek veriyi dilimlemek ve her dilimi farklı bir sınıfa veya değere atamak için kullanılır. [14]

* 1. **Experiment Settings**

Veri ön işlemesi için TF-IDF kullanılmıştır. Sınıflandırma için KNN, Naive Bayes ve Decision Tree kullanılmıştır.

* + 1. **Performance Metrics**

1. **Accuracy:** Toplam doğru tahminlerin toplam tahminlere oranıdır.
2. **Precision:** Pozitif olarak tahmin edilen örneklerin gerçekten pozitif olma oranıdır.
3. **Recall:** Gerçek pozitif örneklerin doğru bir şekilde tahmin edilme oranıdır.
4. **F1 Score:** Kesinlik ve duyarlılığın harmonik ortalamasıdır.[15]

metin, yazı tipi, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 1- Accuracy Formülü

metin, yazı tipi, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 2- Precision Formülü

metin, yazı tipi, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 3- Recall Formülü

metin, yazı tipi, beyaz, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 4 - F1 Score Formülü

1. **Experimental Setting and Result**

Çalışma Python ile ipynb uzantılı olarak visual studio code’da yazılmıştır.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SINIFLANDIRMA | ACCURACY | PRECISION | RECALL | F1 SCORE |
| KNN3EUCLID | 0.78 | 0.84 | 0.78 | 0.77 |
| KNN5EUCLID | 0.76 | 0.83 | 0.76 | 0.74 |
| KNN3MANHATTAN | 0.78 | 0.84 | 0.78 | 0.77 |
| KNN5MANHATTAN | 0.76 | 0.83 | 0.76 | 0.74 |
| DTENTROPY | 0.97 | 0.97 | 0.97 | 0.97 |
| DTGINI | 0.97 | 0.97 | 0.97 | 0.97 |
| NB | 0.92 | 0.92 | 0.92 | 0.92 |

Tablo 2 - Çıktı Tablosu

1. **Conclusion and Discussion**

|  |  |
| --- | --- |
| SINIFLANDIRMA | ACCURACY |
| KNN3EUCLID | 0.78 |
| KNN5EUCLID | 0.76 |
| KNN3MANHATTAN | 0.78 |
| KNN5MANHATTAN | 0.76 |
| DTENTROPY | 0.97 |
| DTGINI | 0.97 |
| NB | 0.92 |

Tablo 3 - Accuracy Sonuçları

Bu çalışmaya göre KNN’nin aldığı parametreler (euclid, manhattan) değişince accuracy’de herhangi bir değişim olmamıştır. KNN3, KNN5’e göre daha iyi bir accuracy verdi. Fakat Decision Tree genel olarak en iyi accuracy’i veren sınıflandırma algoritması olmuştur. Aynı şekilde Decision Tree’nin aldığı parametreler (entropy, gini) değişince accuracy’de herhangi bir değişim olmamıştır.

**Author contributions**

Kaan Berk Duman: Rapor hazırlama.

Ertuğrul Sungur: Sınıflandırma kodları.

Esra Güneş: Veri ön işleme kodları.

İlayda Belen: Rapor için kaynak bilgiler.

**Conflicts of interest**

Yazarlar herhangi bir çıkar çatışması beyan etmemektedir.

**References**

[1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15] <https://chatgpt.com/c/b2af878b-185c-4e38-b82a-76ee8d7932d8>