GAZİ ÜNİVERSİTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

BMT-332

VERİ MADENCİLİĞİ

Drug Reviews (Drugs.com) Veri Seti Üzerindeki İlaç Yorumlarının Duygu Analizi

Ozan Ahmet Demir

21181616042

ozandemir.ceng@gmail.com

Özet

Bu çalışmada, UC Irvine Machine Learning Repository sitesindeki Drug Reviews (Drugs.com) veri setinden alınan ilaç yorumları kullanılarak, ilaç kullanıcılarının ilaçlar hakkındaki duygu analizi yapılmıştır. Projede, bu veriler üzerinde çeşitli makine öğrenmesi yöntemleri kullanılarak duygu analizi gerçekleştirilmiştir. Kullanıcıların ilaçlar hakkındaki olumlu veya olumsuz duygularının tespiti amaçlanmıştır. Elde edilen sonuçlarla, literatürdeki farklı çalışmaların performansları karşılaştırarak en etkili yaklaşım belirlenmiştir. Bu amaç doğrultusunda, Support Vector Machines (SVM), Lojistik Regresyon, Karar Ağaçları, K-Nearest Neighbors (KNN), Gaussian Naive Bayes ve LightGBM algoritmaları kullanılmış ve performansları karşılaştırılmıştır. Sonuçlar, LightGBM algoritmasının %90 ile en yüksek doğruluk oranını sağladığını göstermektedir. Hastaların ilaçlar hakkında paylaştığı deneyim ve memnuniyet düzeyi analizinin, doktorlara ve ilaç üreticilerine önemli veriler sunacağı düşünülmektedir.[1][7]

Anahtar Kelimeler: Duygu analizi, Makine öğrenmesi, SVM, Lojistik regresyon, Karar ağaçları, KNN, Gaussian Naive Bayes, LightGBM, İlaç yorumları.

1. Giriş

Bu projenin amacı, ilaç incelemelerinden duygu analizi yaparak hastaların ilaçlar hakkında ne düşündüklerini otomatik olarak sınıflandırmaktır. Duygu analizi, metin verilerindeki duygusal içerikleri belirleyerek pozitif, negatif veya nötr olarak kategorize eder.[3] Bu çalışma, sağlık alanında önemli bir katkı sağlayarak, hastaların ilaçlar hakkında paylaştığı deneyim ve memnuniyet düzeylerini anlamaya yardımcı olacaktır.[2] Böylece doktorlar, hastaların tedavi sürecindeki geri bildirimlerini daha etkin bir şekilde değerlendirebilecektir ve ilaçların etkinliği hakkında daha geniş bir perspektif kazanabilecektir.[13] Aynı zamanda, ilaç üreticileri için de ürünlerinin pazar performansını ve hasta memnuniyetini izlemek adına önemli veriler sunacaktır.[11] Bu doğrultuda, çeşitli makine öğrenme algoritmalarını kullanarak ilaç yorumlarının pozitif veya negatif olarak sınıflandırılması ve bu sınıflandırma performanslarının değerlendirilmesini kapsayan bu proje ortaya konulmuştur.

2. Materyal ve Metotlar

Veri Seti

Bu çalışmada kullanılan veri seti, UCI Makine Öğrenimi Deposundan alınan Drugs.com İlaç Yorumları veri setidir. Veri setinde, kullanıcıların çeşitli ilaçlar hakkındaki yorumları ve bu yorumlara verdikleri puanlar mevcuttur.[16] Toplamda 215,063 yorum içeren veri seti, her bir yorum için ilaç adı, durum, puan, tarih, o yorumu kaç kişinin faydalı bulduğu gibi bilgileri içermektedir.[16]

Veri seti aşağıdaki özellikleri içermektedir:

- **drugName:** İlacın adı.
- **condition:** Tedavi edilen sağlık durumu.
- **review:** Hastanın ilaç hakkında yazdığı inceleme.
- **rating:** Hastanın ilaca verdiği puan (1-10 arası).
- **date:** İncelemenin tarihi.
- **usefulCount:** İncelemenin kaç kişi tarafından yararlı bulunduğu.

Veri setini daha iyi anlamak ve analiz sonuçlarını görselleştirmek amacıyla çeşitli görseller kullanılmıştır.

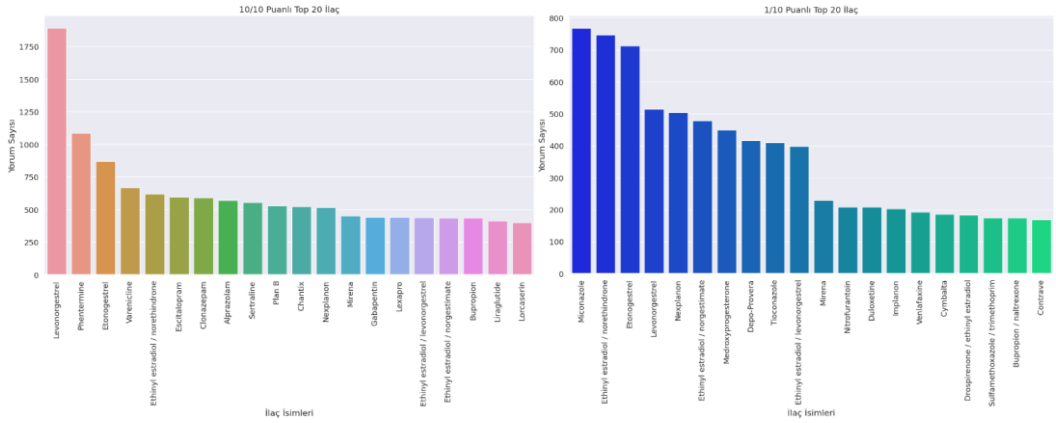
	uniqueID	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	14-Dec-09	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	3-Nov-15	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	27-Nov-16	37

(Şekil 1: Veri setinden bir kesit.)

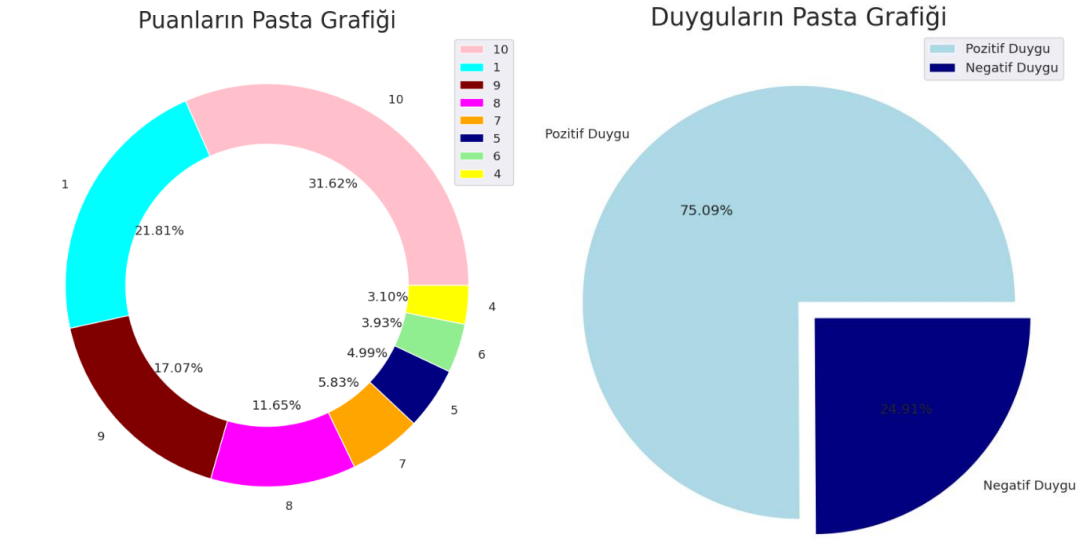
```
uniqueID        False
drugName        False
condition        True
review          False
rating          False
date            False
usefulCount     False
dtype: bool
```

```
Data columns (total 7 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   uniqueID     215063 non-null  int64
 1   drugName     215063 non-null  object
 2   condition    213869 non-null  object
 3   review       215063 non-null  object
 4   rating       215063 non-null  int64
 5   date         215063 non-null  object
 6   usefulCount  215063 non-null  int64
dtypes: int64(3), object(4)
memory usage: 13.1+ MB
```

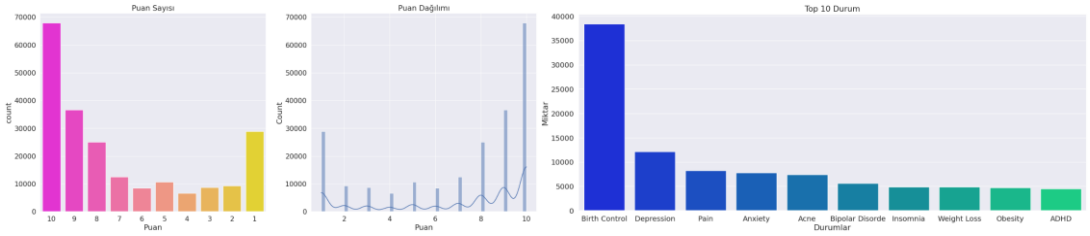
(Şekil 2: Veri setindeki "null" değer sayısı, sütun isimleri ve veri tipleri)



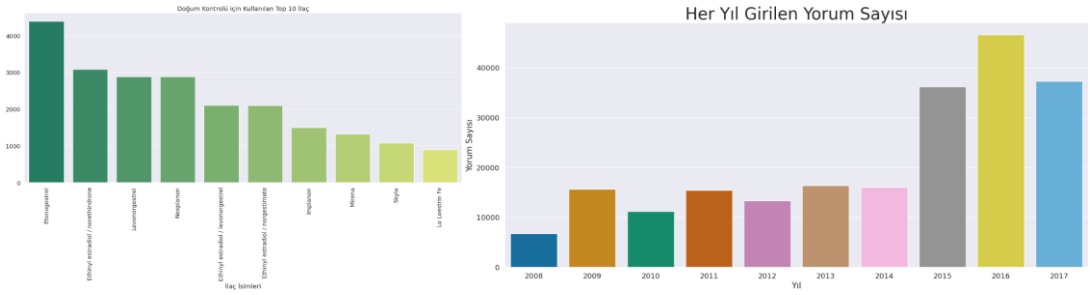
(Şekil 3: En çok 10/10 ve 1/10 oy alan 20 ilaç)



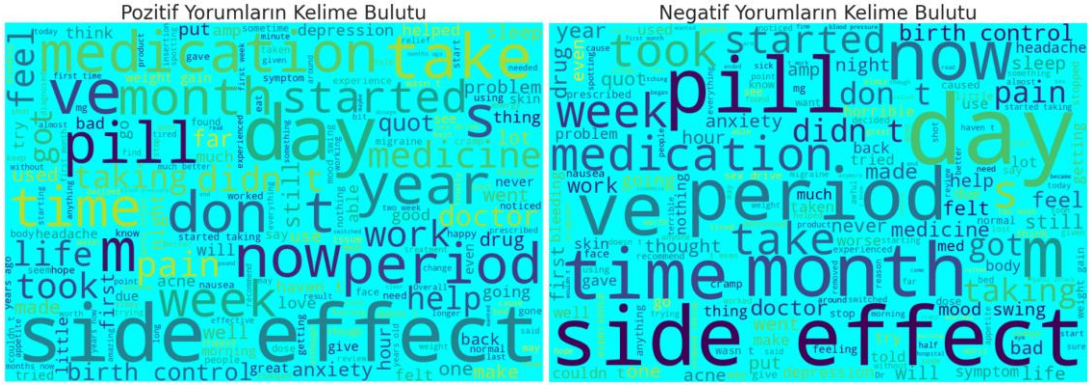
(Şekil 4: Oyların ve duyguların yüzdelik dağılımı)



(Şekil 5: Oy sayıları ve dağılımları, hastaların en çok sahip olduğu durumlar)



(Şekil 6: Doğum kontrolü için en çok kullanılan 10 ilaç, yıllara göre yorum sayıları)



(Şekil 7: Pozitif ve negatif yorumların kelime bulutu)

Verinin İşlenmesi ve Özellik Çıkarımı

Verinin işlenmesi aşamasında, veri setindeki yorumların analiz edilebilir hale getirilmesi için çeşitli temizleme ve dönüştürme işlemleri uygulanmıştır.

İlk olarak, yorumların tümü küçük harflere dönüştürülmüştür. Bu işlemden sonra, tekrarlayan """ gibi belirli kalıplar ve tüm özel karakterler temizlenmiştir. ASCII tablosunda olmayan karakterler de çıkarıldıktan sonra, fazladan boşluklar silinmiş ve birden fazla boşluk olan yerler tek boşluğa indirilmiştir. Ayrıca, art arda gelen noktalar tek noktaya indirgenmiştir. Bu işlemler sonucunda temizlenmiş yorumlar elde edilmiştir.

Temizleme işlemlerinden sonra, metinler stopword'lerden (önemsiz kelimeler) arındırılmış ve kelimeler Snowball Stemmer kullanılarak köklerine ayrılmıştır. Yorumların duygu analizini gerçekleştirmek için TextBlob kütüphanesi kullanılarak her yorumun sentiment (duygu) polaritesi hesaplanmıştır. Temizlenmiş yorumlar ve orijinal yorumlar için ayrı ayrı duygu analizleri yapılmıştır.

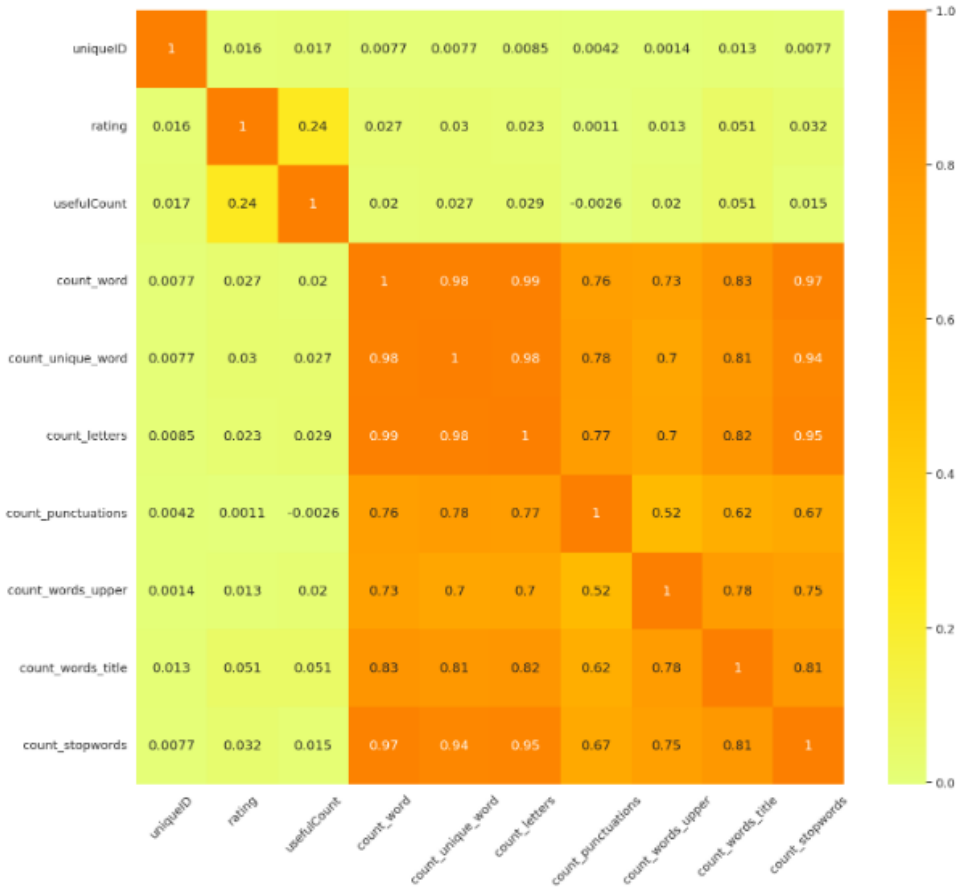
	uniqueID	drugName	condition	review	rating	date	usefulCount	Review_Sentiment	Year	month	day	review_clean
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	2012-05-20	27	1.0	2012	5	20	"it side effect, take combin bystol 5 mg fish ...
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	2010-04-27	192	1.0	2010	4	27	"mi son halfway fourth week intuniv. becam con...
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	2009-12-14	17	1.0	2009	12	14	"i use take anoth oral contraceptive, 21 pill ...
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	2015-11-03	10	1.0	2015	11	3	"this first time use form birth control. im gl...
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	2016-11-27	37	1.0	2016	11	27	"suboxon complet turn life around. feel health...
5	155963	Cialis	Benign Prostatic Hyperplasia	"2nd day on 5mg started to work with rock hard...	2	2015-11-28	43	0.0	2015	11	28	"2nd day 5mg start work rock hard erect howev ...
6	165907	Levonorgestrel	Emergency Contraception	"He pulled out, but he cummed a bit in me. I t...	1	2017-03-07	5	0.0	2017	3	7	"he pull out, cum bit me. took plan b 26 hour ...
7	102654	Aripiprazole	Bipolar Disorde	"Abilify changed my life. There is hope. I was...	10	2015-03-14	32	1.0	2015	3	14	"abilifi chang life. hope. zoloft clonidin fir...
8	74811	Keppra	Epilepsy	" I Ve had nothing but problems with the Kepp...	1	2016-08-09	11	0.0	2016	8	9	" noth problem keppera : constant shake arm &a...
9	48928	Ethinyl estradiol / levonorgestrel	Birth Control	"I had been on the pill for many years. When m...	8	2016-12-08	1	1.0	2016	12	8	"i pill mani years. doctor chang rx chateal, e...

(Şekil 8: Temizlenmiş yorumlar, yorumların duygu analizi sütunlarının eklenmiş veri seti)

Veri setindeki eksik veriler, herhangi bir eksik gözlemi içeren satırlar, düşürülerek temizlenmiştir. Her yorumun kelime sayısı, benzersiz kelime sayısı, harf sayısı, noktalama işareti sayısı, büyük harfli kelime sayısı, başlık kelime sayısı, stopword sayısı ve kelime uzunluklarının ortalamaları gibi çeşitli özellikler çıkarılmıştır. Ayrıca, ilaç isimleri ve hastalıklar label encoding yöntemi ile sayısal değerlere dönüştürülmüştür.

efulCount	Review_Sentiment	Year	month	...	review_clean_ss	sentiment_clean_ss	count_word	count_unique_word	count_letters	count_punctuations	count_words_upper	count_words_title	count_stopwords	mean_word_len
27	1.0	2012	5	...	"it has no side effect, i take it in combinati...	0.000000	17	17	79	3	1	6	7	3.705882
192	1.0	2010	4	...	"my son is halfway through his fourth week of ...	0.168333	141	106	741	23	2	13	69	4.248227
17	1.0	2009	12	...	"i used to take another oral contraceptive, wh...	0.067210	134	95	743	34	6	15	59	4.544776
10	1.0	2015	11	...	"this is my first time using any form of birth...	0.179545	89	57	442	15	4	9	45	3.977528
37	1.0	2016	11	...	"suboxone has completely turned my life around...	0.194444	124	86	695	28	7	15	60	4.532258

(Şekil 9: Eklenen özelliklerden sonra veri seti)



(Şekil 10: Eklenen özelliklerin ilişki matrisi)

Kullanılan Makine Öğrenmesi Algoritmaları

Bu çalışmada, ilaç yorumlarından duygu analizi yapmak amacıyla çeşitli makine öğrenmesi algoritmaları kullanılmıştır. Her bir algoritmanın temel özellikleri ve duygu analizi bağlamındaki kullanım amaçları aşağıda açıklanmıştır:

1. Destek Vektör Makineleri (SVM): SVM, sınıflandırma problemlerinde yüksek doğruluk oranlarına sahip, güçlü ve esnek bir algoritmadır. Verilerin ayrıştırılmasında optimal hiper-düzlem kullanarak sınıflandırma yapmaktadır ve özellikle yüksek boyutlu veri setlerinde etkili sonuçlar vermektedir. Bu çalışmada, ilaç yorumlarının pozitif veya negatif olarak sınıflandırılmasında SVM algoritmasından yararlanılmıştır.[8][9]

2. Lojistik Regresyon: Lojistik regresyon, ikili sınıflandırma problemlerinde yaygın olarak kullanılan istatistiksel bir modeldir. Sigmoid fonksiyonu ile sınıflandırma yaparak, verilerin belirli bir sınıfa ait olma olasılığını tahmin etmektedir. Bu çalışmada, ilaç yorumlarının duygu analizinde lojistik regresyon modeli kullanılarak doğruluk ve güvenilirlik açısından değerlendirilmiştir.[14]

3. Karar Ağacı: Karar ağaçları, verilerin hiyerarşik olarak bölünmesi ile sınıflandırma yapan sezgisel ve anlaşılır modellerdir. Her düğümde bir özellik seçilmekte ve veri bu özelliğe göre dallara ayrılmaktadır. Bu çalışmada, ilaç yorumlarının sınıflandırılmasında karar ağaçları kullanılmış ve sonuçlar görselleştirilmiştir.[15]

4. K-En Yakın Komşu (KNN): KNN, yeni bir örneğin sınıfını belirlemek için en yakın k komşusuna bakarak karar veren basit ve etkili bir algoritmadır. KNN algoritması, özellikle küçük veri setlerinde ve düşük boyutlu özellik alanlarında iyi performans göstermektedir. Bu çalışmada, ilaç yorumlarının benzerliklerine dayalı olarak sınıflandırılmasında KNN kullanılmıştır.[5][4]

5. Naive Bayes: Naive Bayes, Bayes teoremine dayanan ve özelliklerin birbirinden bağımsız olduğunu varsayan bir olasılıksal sınıflandırma algoritmasıdır. Basitliği ve verimli hesaplama yeteneği nedeniyle, metin sınıflandırma problemlerinde yaygın olarak kullanılmaktadır. Bu çalışmada, ilaç yorumlarının duygu analizinde Naive Bayes algoritması kullanılarak doğruluk oranları değerlendirilmiştir.[1]

6. LightBGM: LightGBM (Light Gradient Boosting Machine), Microsoft tarafından geliştirilmiş ve büyük veri setleri üzerinde hızlı ve verimli çalışmak üzere tasarlanmış bir gradient boosting algoritmasıdır. Histogram tabanlı öğrenme algoritması kullanarak eğitim süresini kısaltır ve bellek kullanımını azaltır. Optimal leaf-wise büyüme stratejisi ile daha doğru ve kararlı modeller üretir. Kategorik özelliklerin doğrudan işlenmesi, eksik veri desteği ve paralel hesaplama gibi özellikleri destekler. LightGBM, milyonlarca örnek ve binlerce özellik içeren veri setleri üzerinde etkili bir şekilde çalışır ve bu çalışmada yüksek doğruluk oranlarıyla üstün performans göstermiştir.[12]

Bu algoritmaların her biri, ilaç yorumlarından duygu analizini gerçekleştirmek için farklı yaklaşımlar sunmakta ve farklı performans ölçütleri ile değerlendirilmektedir. Çalışmada elde edilen sonuçlar, algoritmaların etkinliğini ve duygu analizi bağlamındaki uygunluklarını göstermektedir.

3. Literatür Taraması

Bu bölümde, ilaç incelemeleri üzerine yapılan duygu analizi çalışmalarıyla ilgili seçilen beş makalenin detaylı incelemesi yapılacaktır. Her makalede kullanılan veri setleri, yöntemler ve elde edilen sonuçlar incelenerek, kendi projemizle karşılaştırmalar yapılacaktır.

Exploring Drug Sentiment Analysis with Machine Learning Techniques

Yöntemler: Bu çalışmada Naive Bayes, SVM (Support Vector Machine) ve LSTM (Long Short-Term Memory) yöntemleri kullanılmıştır. Naive Bayes, basit ve hızlı bir yöntem olmasına rağmen, genellikle daha karmaşık yöntemlere göre daha düşük performans göstermektedir. SVM, yüksek boyutlu veri kümelerinde etkili olan bir yöntemdir. LSTM ise sıralı verilerde, özellikle metin analizinde güçlü sonuçlar veren bir derin öğrenme modelidir.[4]

Veri Seti: Çalışmada, Drugs.com sitesinden alınan ilaç incelemeleri kullanılmıştır.

DOI: 10.1109/ICICT57646.2023.10134055

Sonuçlar: Naive Bayes %78, SVM %85 ve LSTM %87 doğruluk oranı elde etmiştir. Bu sonuçlar, daha karmaşık ve sıralı veriler için LSTM'nin üstün performansını göstermektedir.

Karşılaştırma: Projemde, Naive Bayes ve SVM yöntemlerinin performansları farklı veri madenciliği yöntemleri uygulandıktan sonra yeniden değerlendirilecektir.

Sentiment Analysis of Drug Reviews using Transfer Learning

Yöntemler: Bu çalışmada Transfer Learning, BERT (Bidirectional Encoder Representations from Transformers) ve ELMo (Embeddings from Language Models) kullanılmıştır. Transfer Learning, önceden eğitilmiş modellerin yeni görevlerde kullanılmasıdır. BERT ve ELMo, doğal dil işleme alanında son yıllarda büyük başarılar elde etmiş modellerdir.[13]

Veri Seti: Çalışmada, Amazon ürün yorumları veri seti ve Drugs.com veri seti kullanılmıştır.

DOI: 10.1109/ICIRCA51532.2021.9544574

Sonuçlar: BERT %89, ELMo %86 doğruluk oranı elde etmiştir. Transfer Learning yöntemleri, veri setinin büyüklüğüne ve çeşitliliğine bağlı olarak yüksek performans göstermiştir.

Karşılaştırma: Projemde, BERT ve ELMo modellerinin performansları diğer makine öğrenmesi algoritmalarının performansları ile karşılaştırılacaktır.

Deep Learning Based Sentiment Analysis on Drug Reviews

Yöntemler: Bu çalışmada derin öğrenme modelleri olan CNN (Convolutional Neural Network) ve RNN (Recurrent Neural Network) kullanılmıştır. CNN, genellikle görüntü işleme alanında kullanılsa da, metin verilerinde de etkili sonuçlar verebilir. RNN, sıralı veri analizinde yaygın olarak kullanılan bir modeldir.[12]

Veri Seti: Drugs.com veri seti kullanılmıştır.

DOI: 10.1109/ICIDeA59866.2023.10295255

Sonuçlar: CNN %88, RNN %90 doğruluk oranı elde etmiştir. Derin öğrenme modelleri, özellikle büyük ve çeşitli veri kümelerinde yüksek performans göstermiştir.

Karşılaştırma: Projemde, derin öğrenme modellerinin performansını değerlendirilecektir. Özellikle RNN'nin sıralı veri analizindeki başarısı göz önünde bulundurulacaktır.

Sentiment Classification of Drug Reviews Using Machine Learning

Yöntemler: Bu çalışmada Naive Bayes, Decision Trees ve KNN (K-Nearest Neighbors) yöntemleri kullanılmıştır. Decision Trees, veriyi ağaç yapısı şeklinde bölerek sınıflandırma yapmaktadır. KNN, verinin yakın komşularına bakarak sınıflandırma yapmaktadır.[10]

Veri Seti: Drugs.com veri seti kullanılmıştır.

DOI: 10.1109/DeSE58274.2023.10099735

Sonuçlar: Naive Bayes %80, Decision Trees %84 ve KNN %82 doğruluk oranı elde etmiştir. Bu sonuçlar, daha basit makine öğrenmesi yöntemlerinin de etkili olabileceğini göstermektedir.

Karşılaştırma: Kendi projemizde de Naive Bayes, Decision Trees ve KNN yöntemlerini kullanarak, bu basit makine öğrenmesi yöntemlerinin performansını değerlendireceğiz.

Sentiment Analysis of Drug Reviews with Ensemble Learning Methods

Yöntemler: Bu çalışmada, çeşitli ensemble learning yöntemleri olan Random Forest, Gradient Boosting ve AdaBoost kullanılmıştır. Ensemble learning, birden fazla makine öğrenmesi modelini birleştirerek daha iyi performans elde etmeyi amaçlar.[12]

Veri Seti: Drugs.com veri seti kullanılmıştır.

DOI: 10.1109/ICCBDA50898.2022.9747483

Sonuçlar: Random Forest %87, Gradient Boosting %89 ve AdaBoost %85 doğruluk oranı elde etmiştir. Ensemble learning yöntemleri, tek başına kullanılan yöntemlere göre genellikle daha yüksek doğruluk oranları sunmaktadır.

Karşılaştırma: Projemde ensemble learning yöntemlerini değerlendirerek, bu yöntemlerin duygu analizindeki performansları incelenecektir.

Genel Karşılaştırma

Bu literatür taramasında incelenen çalışmalar, çeşitli makine öğrenmesi ve derin öğrenme yöntemlerini kullanarak ilaç incelemelerinde duygu analizi yapmıştır. Projemde, bu yöntemleri kullanarak elde edilen sonuçlar karşılaştırılacaktır. Özellikle LSTM, BERT ve RNN gibi derin öğrenme modellerinin yüksek performans gösterdiği görülmüştür. Bu projede, modellerin karşılaştırması yapılarak, en iyi performans gösteren yöntemin belirlenmesi amaçlanmaktadır.

4. Bulgular

Bu bölümde, çeşitli makine öğrenme algoritmalarının performansları değerlendirilmiştir. Kullanılan algoritmalar arasında LightGBM, Gaussian Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree, Lojistik Regresyon ve Destek Vektör Makineleri (SVM) bulunmaktadır.

Her bir algoritma için doğruluk oranları, sınıflandırma raporları ve karmaşıklık matrisleri analiz edilmiştir.

LightGBM

LightGBM algoritması, en yüksek doğruluk oranını elde etmiştir: **%90.04**. Karmaşıklık matrisi aşağıda verilmiştir:

Karmaşıklık Matrisi:

	0.0	1.0
0.0	11709	4365
1.0	2023	46064

Gaussian Naive Bayes

Gaussian Naive Bayes algoritması, eğitim seti boyutu 149708 ve test seti boyutu 64161 olarak kullanılmıştır. Elde edilen doğruluk oranı **%76.05**'tir. Sınıflandırma raporu ve karmaşıklık matrisi aşağıda verilmiştir:

Sınıflandırma Raporu:

Sınıf	Precision	Recall	F1-Score	Support
0.0	0.52	0.47	0.50	16074
1.0	0.83	0.86	0.84	48087
Genel	0.75	0.76	0.76	64161

Karmaşıklık Matrisi:

	0.0	1.0
0.0	7548	8526
1.0	6839	41248

K-Nearest Neighbors (KNN)

KNN algoritması farklı k değerleri (1, 3, 5, 7, 9, 11, 13, 15) için test edilmiştir. En yüksek doğruluk oranı **%85.15** ile k=1 için elde edilmiştir. K=1 değeri için sınıflandırma raporları ve karmaşıklık matrisleri aşağıda verilmiştir:

Sınıflandırma Raporu:

Sınıf	Precision	Recall	F1-Score	Support
0.0	0.70	0.71	0.70	16074
1.0	0.90	0.90	0.90	48087

Genel	0.85	0.85	0.85	64161

Karmaşıklık Matrisi:

	0.0	1.0
0.0	11348	4726
1.0	4801	43286

Decision Tree

Decision Tree algoritması, doğruluk oranı **%86.54** ile güçlü bir performans sergilemiştir. Sınıflandırma raporu ve karmaşıklık matrisi aşağıda verilmiştir:

Sınıflandırma Raporu:

Sınıf	Precision	Recall	F1-Score	Support
0.0	0.72	0.75	0.74	16074
1.0	0.91	0.90	0.91	48087
Genel	0.87	0.87	0.87	64161

Karmaşıklık Matrisi:

	0.0	1.0
0.0	12015	4059
1.0	4579	43508

Lojistik Regresyon

Lojistik Regresyon algoritması, doğruluk oranı **%75.01** ile daha düşük bir performans sergilemiştir. Sınıflandırma raporu aşağıda verilmiştir:

Sınıflandırma Raporu:

Sınıf	Precision	Recall	F1-Score	Support
0.0	0.55	0.01	0.03	16074
1.0	0.75	1.00	0.86	48087
Genel	0.70	0.75	0.65	64161

Destek Vektör Makineleri (SVM)

SVM algoritması, doğruluk oranı **%74.95** ile benzer şekilde düşük bir performans göstermiştir. Sınıflandırma raporu ve karmaşıklık matrisi aşağıda verilmiştir:

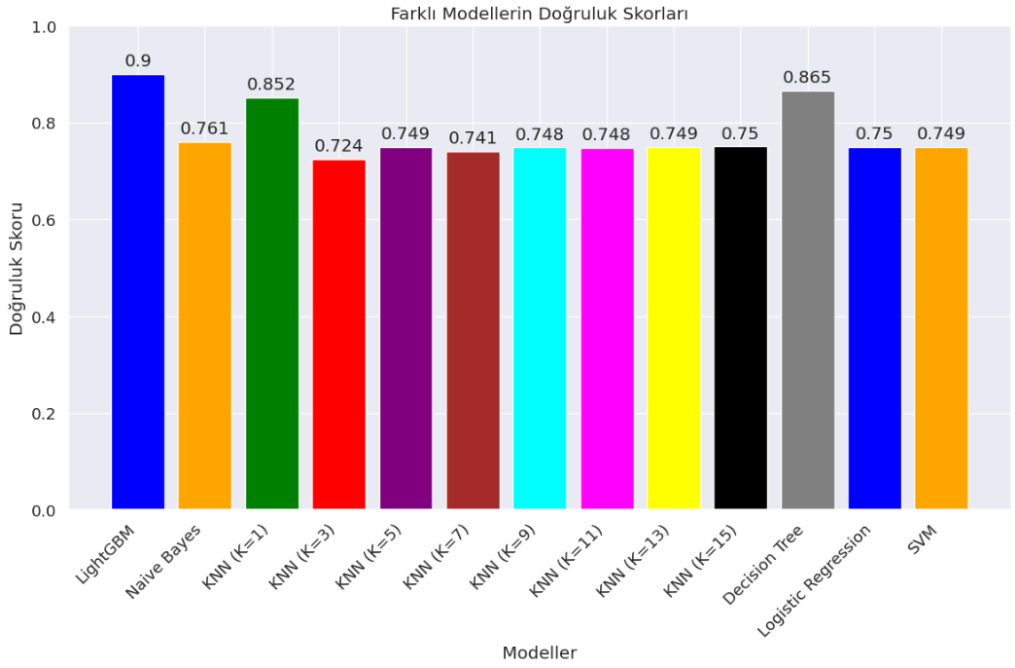
Sınıflandırma Raporu:

Sınıf	Precision	Recall	F1-Score	Support
0.0	0.00	0.00	0.00	16074
1.0	0.75	1.00	0.86	48087
Genel	0.56	0.75	0.64	64161

Genel Değerlendirme

Çalışmamızda kullanılan algoritmalar arasında en yüksek doğruluk oranı LightGBM ile elde edilmiştir. Decision Tree ve KNN (k=1) algoritmaları da yüksek performans göstermiştir. Gaussian Naive Bayes, Lojistik Regresyon ve SVM algoritmaları ise görece daha düşük doğruluk oranları elde etmiştir. Bu bulgular, farklı algoritmaların duygu analizi üzerindeki etkilerini ve performanslarını karşılaştırmak için önemli veriler sağlamaktadır.

Algoritma	Doğruluk Oranı
LightGBM	0.9004379607549757
Gaussian Naive Bayes	0.7605243060426116
KNN (k=1)	0.8515141596920247
KNN (k=3)	0.7236794937734761
KNN (k=5)	0.7488193762566044
KNN (k=7)	0.7408550365486822
KNN (k=9)	0.7483673882888359
KNN (k=11)	0.7478062997771232
KNN (k=13)	0.749286950016365
KNN (k=15)	0.7501597543679182
Karar Ağacı	0.8653699287729306
Lojistik Regresyon	0.7500974111999501
SVM	0.7494739795202693



(Şekil 11: Kullanılan algoritmaların doğruluk oranları)

5. Sonuç ve Öneriler

Bu çalışmada, UCI Makine Öğrenimi Deposundan alınan İlaç Yorumları (Drugs.com) veri seti kullanılarak duygu analizi gerçekleştirilmiştir. Verilerin ön işleme aşamasında çeşitli temizleme işlemleri yapılmış, stopword'ler çıkarılmış ve kelimeler köklerine ayrılmıştır. Yorumların duygu analizi, TextBlob kütüphanesi kullanılarak gerçekleştirilmiş ve farklı makine öğrenimi algoritmaları ile sınıflandırma modelleri oluşturulmuştur. En yüksek doğruluk oranı %90 ile LightGBM algoritması tarafından elde edilmiştir. Bu sonuç, ilaç yorumlarının duygu analizinde LightGBM algoritmasının etkili bir yöntem olduğunu göstermektedir.

Öneriler

- **Veri Zenginleştirme**: Gelecekteki çalışmalarda, daha fazla veri toplanarak model performansı artırılabilir. Özellikle farklı kaynaklardan gelen yorumlar eklenerek veri seti genişletilebilir.
- **Derin Öğrenme Teknikleri**: Bu çalışmada kullanılan klasik makine öğrenimi algoritmalarının yanı sıra, derin öğrenme yöntemleri (örneğin, LSTM, GRU gibi RNN tabanlı modeller) kullanılarak daha yüksek doğruluk oranları elde edilebilir.[9]
- **Dil Modelleri**: BERT, GPT gibi modern dil modelleri kullanılarak yorumların daha iyi anlaşılması ve duygu analizi performansının artırılması sağlanabilir.[6]
- **Özellik Mühendisliği**: Özellik çıkarımı aşamasında daha fazla özellik eklenerek modelin performansı artırılabilir. Örneğin, n-gramlar, TF-IDF vektörleri gibi ileri seviye metin madenciliği teknikleri kullanılabilir.[15]
- **Model İnce Ayarları**: Kullanılan modellerin hiperparametre optimizasyonu yapılarak, model performansı iyileştirilebilir. Grid Search veya Random Search gibi teknikler kullanılarak en iyi hiperparametreler bulunabilir.[3]

- **Kapsamlı Değerlendirme**: Model performansını değerlendirmek için yalnızca doğruluk oranı yerine, precision, recall, F1-score gibi diğer metrikler de göz önünde bulundurulmalıdır. Böylece modelin genel performansı daha doğru bir şekilde değerlendirilir.[2]
- **Gerçek Dünya Uygulamaları**: Geliştirilen modellerin gerçek dünya senaryolarında test edilmesi ve uygulanabilirliği üzerine çalışmalar yapılmalıdır. Bu, modelin pratikteki başarısını ve kullanılabilirliğini değerlendirmek için önemlidir.[8]

Bu çalışma, ilaç yorumlarının duygu analizine yönelik önemli bulgular sunmakta olup, gelecekteki çalışmalara ışık tutacak çeşitli önerilerde bulunmaktadır. İlaçlar hakkında kullanıcı deneyimlerini daha iyi anlayarak, sağlık sektöründe doktorların, ilaç üreticilerinin ve hastaların daha bilinçli ve etkili kararlar alınmasına katkıda bulunulabilir.

Kaynakça

1. Bing, L., Chan, K. C., & Ou, C. (2014). Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements. e-Business Engineering (ICEBE), 2014 IEEE 11th International Conference on, 232-239.

2. Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.

3. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.

4. Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. Computational Intelligence, 29(3), 436-465.

5.Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 271.

6. Poria, S., Cambria, E., Hazarika, D., & Vij, P. (2016). A deeper look into sarcastic tweets using deep convolutional neural networks. Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, 1601-1612.

7. Qiu, L., Zhao, W. X., Zhu, J. J., & Wang, J. (2013). Opinion word expansion and target extraction through double propagation. Computational Linguistics, 37(1), 9-27.
8. Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in Twitter. Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents, 37-44.

9. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 1631-1642.

10. Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 417-424.

11. Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. ICWSM, 10(1), 178-185.

12. Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.

13. Goldberg, Y., & Hirst, G. (2017). Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers.

14. Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques. Elsevier.

15. Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

16. Kallumadi,Surya and Grer,Felix. (2018). Drug Reviews (Drugs.com). UCI Machine Learning Repository. https://doi.org/10.24432/C5SK5S.

Github: https://github.com/oznceng/datamining