Doğal Dil İşleme Modelleri Kullanarak Ürün Yorumlarının Olumlu, Olumsuz ve Nötr Durumlarının Tahmin Edilmesi

Muhammet Semih KELEŞ
İzmir Kâtip Çelebi Üniversitesi
İzmir, Türkiye
semihkeles1997@hotmail.com

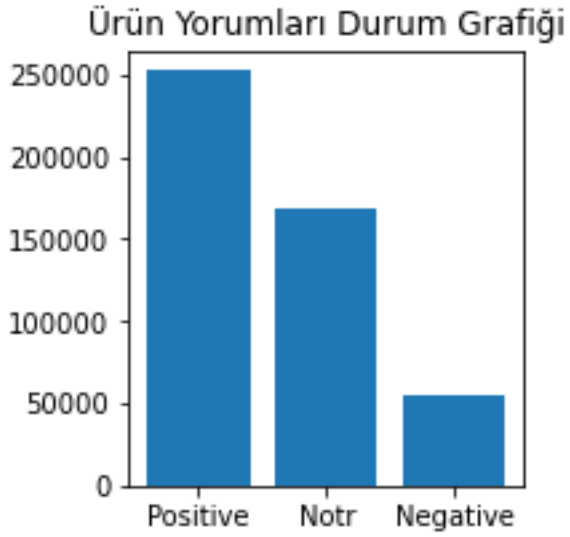
I. GİRİŞ (INTRODUCTION)

İnsan var olduğu süre boyunca mevcut ihtiyaçlarını karşılamak durumundadır. Bu ihtiyaçlar kimi zaman takas yöntemi ile olurken kimi zaman nakit para üzerinden gerçekleşmektedir. Zaman geçtikçe insanların ihtiyaçları da sürekli olarak gelişmektedir. İhtiyaçlar arttıkça da ilgili ihtiyaca çözüm üretebilecek birçok ürün ortaya çıkmaktadır. Hâl böyle iken hangi ürünün ilgili ihtiyaçlara yönelik olduğunu anlamak, günümüzde oldukça popüler bir söz öbeği haline gelmiş "fiyat-performans ürünü" bulabilmek bir hâyli zor olabilmektedir. Bununla birlikte insanların ihtiyaçları dışında alşveriş yaptığı da görülmüştür. İnsanların tarih boyunca bir şeyler satın almak için moda, popüler olma ve taklit etme gibi çeşitli motivasyon kaynakları olmuştur. ([1]) [2] Dolayısıyla alışveriş yapmak insanlar için vazgeçilmez ve sürekli gerçekleşen bir durum olarak görülmektedir. İlgili ürünlerin bulunabilmesi için çevrimiçi ürün satış platformları, çevrimiçi sosyal medya platformları gibi birçok alan ortaya çıkmıştır. Bu durum ise alışveriş yapmak için daha az efor harcanması konusunda fayda sağlamıştır. Böylece insanlar mağazaya gitmeden de alışveriş yapabilir hâle gelmişlerdir. Ancak bu durumun getirdiği bazı eksiklikler de olmuştur. Bu eksiklikler ürün hakkında bilgi sahibi olmak için ürünün satıldığı platformdaki açıklamalarla yetinmek durumunda kalınmasıdır. Birçok platform ürün bilgisi, mağazaya soru sorma imkânları tanısa da insanlar satıcıya tüketici kadar güvenmemektedir. Bu durumda ürün yorumları oldukça önem arz etmektedir. İnsanlar kendileri gibi çevrimiçi ortamda alışveriş yapan insanların satın alınacak ürün hakkında yorum yapmasını önemsemektedir. Bu durum da ürün yorumlarının ne kadar önemli olduğunu ortaya koyabilmektedir. Bu çalışma da ürün yorumlarının pozitif, negatif veya nötr olma durumunun tahmin edilmesi üzerine gerçekleşmektedir. İlgili veri seti 489644 veriden oluşmaktadır. İlgili veri seti Türkçe durak kelimelerinden (stopwords), linklerden, özel karakterlerden, sayısal ifadelerden ve anlamsız kelimelerden arındırılmıştır. Daha sonra Lemmatizer ve PorterStemmer işlemleri uygulanmış; Lemma'nın daha performanslı çalıştığı görüldüğünden lemma ile devam edilmiştir. Daha sonra ilgili veri seti Naive Bayes ve KNN algoritmaları kullanılarak aşağıda belirtilen 6 farklı modelle eğitilmiştir.

II. YÖNTEMLER (METHODS)

A. *Veri Seti (Data Set)*

Bu çalışmada kullanılan veri seti çeşitli elektronik mağazalarından toplanmıştır. İlgili veri setinde ürün yorumları ve duygu durumları öznitelikleri olmak üzere 2 öznitelik ve 489644 örnek bulunmaktadır. Bu örneklerden 262166 tanesi pozitif, 170917 tanesi nötr ve 56561 tanesi ise negatif yorumlardan oluşmaktadır. (Grafik v1)



Grafik v1

B. *Metin Sınıflandırma (Text Classification)*

Metin sınıflandırma işlemleri için verilerin kategorize edilmesi ve etiketlenmesi gerekmektedir. Bununla birlikte veriler doğrudan kullanılamamakta ve belli başlı ön işlemlerden geçirilmesi gerekmektedir. Bu işlemlerden biri de öznitelik vektörüne dönüştürmedir. Metin sınıflandırma işlemlerinde öznitelik vektörüne dönüştürme başarısı sınıflandırma başarısını doğrudan etkilemektedir. [3]

C. *Ön işleme (Preprocessing)*

1) Kök Bulma (Stemming)

Kök bulma işlemi de Metin Sınıflandırma işlemleri için gereken ön işlemlerden bir tanesidir. Bu işlemde metinde geçen kelimelerin kökleri bulunmaktadır. Bu sayede aynı köke sahip kelimeler 1 kelime olarak sayılabilmekte ve her kelimenin bir özniteliği temsil ettiği düşünüldüğünde öznitelik sayısında önemli ölcüde azalma görülebilecektir. Bu yöntem özellikle Türkçe gibi sondan eklemeli dillerde sıklıkla tercih edilmektedir.

2) Durdurma Kelimelerini Kaldırma (Stopword Removing)

Bu aşamada genellikle tek başlarına bir anlam ifade etmeyen ancak çok sık kullanılan kelimeler (ve, veya, ile, ama, acaba, aslında, bir, gibi ...) kaldırılmaktadır. Bu sayede öznitelik vektör boyutu gereksiz yere büyümeyecektir.

3) Terim Frekansı - Ters Doküman Frekansı (Term Frequency - Inverse Document Frequency)

Terim Frekansı (Term Frequency-TF) metinde geçen her bir kelimenin metindeki kullanma sıklığını hesaplamak için kullanılan ön işlemdir. Formülü aşağıdaki gibidir.

$$\text{TF} = \frac{\text{Terimin dokümanda geçme adedi}}{\text{Dokümandaki toplam terim adedi}}$$

Ters Doküman Frekansı (Inverse Dokument Frequency-IDF) ise bir terimin arandığı tüm doküman sayısının o terimin bulunduğu doküman sayısına oranını vermektedir.

$$\text{IDF} = \frac{\text{Toplam doküman sayısı}}{\text{Terimin geçtiği doküman sayısı}}$$

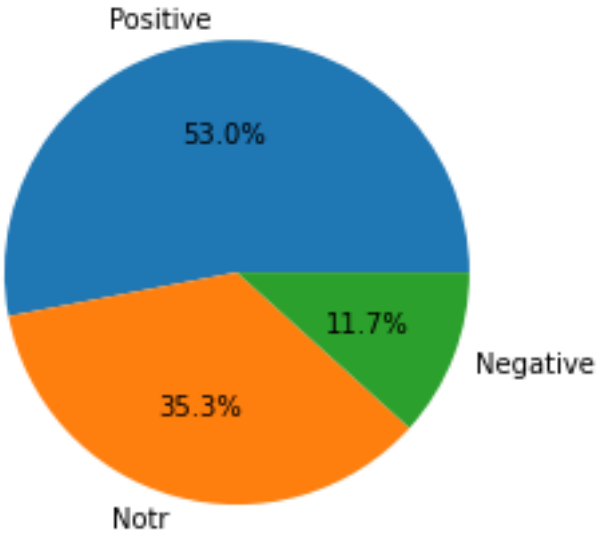
TF-IDF ise Terim Frekansı ile Ters Doküman Frekansının çarpımı ile elde edilmektedir. Bu yöntem ile birlikte bir terimin bulunduğu dokümanın sınıflandırılmasında katkı sağlamadaki önemini göstermektedir.

$$\text{TF-IDF} = \text{TF} * \text{IDF}$$

4) N-Grams (N-Grams)

Metin önişleme aşamalarından biri olan N-Gram ise metinde geçen kelimelerin gruplandırılarak kullanılmasıdır. 1 kelimelik gruplar unigram, 2 kelimelik gruplar bigram, 3 kelimelik gruplar trigram olarak adlandırılmaktadır. N-Gram işlemi 3'ten fazla olan n sayıdaki grubu temsil etmektedir. Örneğin; "bugün hava çok sıcak" cümlesi için bigram olarak bölünmek istenirse ["bugün hava","hava çok","çok sıcak"] vektörü elde edilecektir.

İlgili çalışmada kök bulma işlemi, durak kelimelerinin kaldırılması ve TF-IDF işlemleri yapılmıştır. Yapılan işlemler sonucunda 6 farklı modelle eğitilmiştir. Veri ön işleme aşamasında lemmatizer işlemleri yapılmış, tek harflik cümleler, boş cümleler ve tekrar eden cümleler kaldırılmıştır. Bu işlemler sonucunda veri seti 252196 pozitif, 167880 nötr ve 55561 negatif olmak üzere 475637 örneğe düşürülmüştür. İlgili örneklerin



Grafik v2

MODEL - 1	
Vectorizer	TF-IDF
Algoritma	Naive Bayes
Max Features	-
N-gram	-
Binary	-

MODEL - 2	
Vectorizer	Count Vectorizer
Algoritma	Naive Bayes
Max Features	1000
N-gram Range	-
Binary	True

MODEL - 3	
Vectorizer	Count Vectorizer
Algoritma	Naive Bayes
Max Features	1000
N-gram Range	(1,2)
Binary	-

MODEL - 4	
Vectorizer	Count Vectorizer
Algoritma	Naive Bayes
Max Features	1000
N-gram Range	-
Binary	-

MODEL - 5	
Vectorizer	TF-IDF
Algoritma	Naive Bayes
Max Features	1000
N-gram Range	(1,2)
Binary	-

MODEL - 6	
Vectorizer	TF-IDF
Algoritma	KNN
Max Features	1000
N-gram Range	(1,2)
Binary	-

D. Deneylerde Kullanılan Algoritmalar

K-En Yakın Komşu Algoritması (K-Nearest Neighbour KNN Algorithm)

Sınıflandırma ve regresyon işlemlerinde kullanılabilen ve sıklıkla tercih edilen algoritmalardan biridir. Özniteliklerin birbirlerine olan uzaklıklarının Öklid işlemi ile hesaplanması ile isminde de geçen K parametresi ile en yakın K sayıdaki örneğe bakarak sınıflandırma işleminin yapıldığı algoritmadır. [4] [5] İşlemin sonunda örneğimiz, K parametresi kadar örneğin içerisinden çoğunlukla hangi sınıfa giriyor ise KNN algoritması çoğunluğun bulunduğu etiketi vermektedir. Örneğin K parametresi 3 verilmiş ve işlemin sonunda ulaşılmış örneklerin sınıf etiketleri "Başarılı","Başarılı" ve "Başarısız" ise KNN algoritması etiketlemek istediğimiz örneği "Başarılı" etiketi ile etiketleyecektir. Özellikle sınıflandırma işlemlerinde genellikle K değeri tek sayı olarak seçilmektedir. Bunun sebebi işlemin sonunda elde edilecek etiketlerin yarı yarıya olma durumunu engellemektir. Örneğin K değerinin 4 seçilme durumunda ve işlemin sonunda "Başarılı", "Başarılı", "Başarısız" ve "Başarısız" etiketlerine sahip örneklerin getirilmesi sonucunda KNN algoritması sağlıklı bir etiketleme işlemi yapamayacaktır. Regresyon işleminde de aynı işlemler yapılıp K kadar örneğin ortalaması hesaplanmaktadır.

Naive Bayes Algoritması

Genellikle metin sınıflandırma projelerinde kullanılan Naive Bayes algoritması supervised (denetimli) öğrenme modeli ile gerçekleşen bir sınıflandırma algoritmasıdır. Bağımsız değişkenin, bağımlı değişken üzerindeki etkilerine göre istatistik ve olasılık temellerine göre yeni bir durumu sınıflandırma işlemidir.
Durumların olasılık değerlerinin hesaplandığı formül:

$$X = [x_1, x_2, x_3, ...x_n]$$

$$C = [c_1, c_2, c_3, ...c_n])$$

$$P(C_i/x) = \frac{p(x/C_i)}{P(x)}$$

Yeni durum ise $argmax_{ci} P(x/c_i)P(c_i)$ formülü kullanılarak hesaplanan en yüksek olasılık değerinin sınıfına atanır. [6] [5]

E. Öznitelik Seçimi (Feature Selection)

Öznitelik seçimi, kullanılan veri kümesinin özniteliklerinden sınıflandırma başarısına en fazla katkı sağlayanların tespit edilip seçilmesi işlemidir. Öznitelik seçimindeki amaç, sınıflandırma başarısını yükseltmek veya eğitim süresini kısaltarak çalışma performansını arttırmaktır. Metin veri kümelerinin öznitelik sayısının yüksek olması sebebiyle, metin sınıflandırmada kritik bir önişlem olarak sıklıkla kullanılmaktadır. Öznitelik seçimi işleminde; filtreleme, sarmalama ve gömülü yöntemler kategorilerinde çeşitli algoritmalar bulunmaktadır. Bunlardan öznitelik ve sınıf değişken tiplerine uygun olacak şekilde en yaygın kullanılan yöntemler; Pearson korelasyonu, Ki-kare testi, Anova testi ve Bilgi kazanımı yöntemleridir. [3]

III. SINIFLANDIRMA ÖLÇÜTLERI (CLASSIFICATION METRICS)

Sınıflandırma modellerinde başarıyı ölçümleyebilmek için belli başlı sınıflandırma ölçütleri kullanılmaktadır. Genellikle doğruluk (Accuracy), kesinlik (Precision), duyarlılık (Recall) ve F-1 skorları değerlendirilir.

1) Karışıklık Matrisi (Confusion Matrix)

Karışıklık Matrisi, sınıflandırma işlemindeki doğru ve yanlış tahminlerinin sayısal ifadesini vermektedir.

TP	True Positive	Pozitif değeri doğru bir şekilde etiketlemek
TN	True Negative	Negatif değeri doğru bir şekilde etiketlemek
FN	False Negative	Negatif değeri hatalı şekilde etiketlemek
FP	False Positive	Pozitif değeri hatalı şekilde etiketlemek

Örneğin olumlu bir yorumu "olumlu" şeklinde etiketlemek TP örneğidir. Olumlu bir yorumu "olumsuz" şeklinde etiketlemek FP örneğidir. Olumsuz bir yorumu "olumsuz" şeklinde etiketlemek TN örneği iken olumsuz bir yorumu "olumlu" şeklinde etiketlemek ise FN örneğidir.

*2) **Doğruluk (Accuracy)***

Doğruluk, sınıflandırma işlemi sonucundaki bütün tahminlerdeki doğru tahmin oranı olarak tanımlanmaktadır. [3]

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

*3) **Kesinlik (Precision)***

Kesinlik, sınıflandırma işlemi sonucundaki "Pozitif" olarak tahmin edilenlerin hangi oranda gerçekten "Pozitif" sınıfına ait olduğunu gösteren sınıflandırma ölçütüdür. Kesinlik değerinin matematiksel hesabı aşağıdaki eşitlikte görülmektedir. [3]

$$\text{Precision} = \frac{TP}{TP+FP}$$

*4) **Duyarlılık (Recall)***

Duyarlılık, "Pozitif" sınıfındaki örneklerin hangi oranda "Pozitif" olarak tahmin edildiğini gösteren sınıflandırma ölçütüdür. Duyarlılık değerinin matematiksel hesabı aşağıdaki eşitlikte görülmektedir. [3]

$$\text{Recall} = \frac{TP}{TP+FN}$$

*5) **F-1 Skoru (F-1 Score)***

F1-Skor ölçütü değeri, kesinlik ve duyarlılık değerlerinin harmonik ortalaması sonucu elde edilmektedir. F1-Skor değerinin matematiksel hesabı aşağıdaki eşitlikte görülmektedir. [3]

$$\text{F-1 Score} = 2 * \frac{Kesinlik * Duyarlilik}{Kesinlik + Duyarlilik}$$

IV. Sonuçlar ve Tartışma

İlgili çalışma 489644 örnek ile başlamış ve önişleme işlemleri sonucunda 475637 örneğe düşürülmüştür. Bu örneklerin yaklaşık %10'u (48965 örnek) test veri seti ve yaklaşık %90'ı (440679 örnek) ise eğitim veri seti olarak ayrılmıştır. İlgili eğitim veri seti 6 farklı modelle MorphAnalyzer işlemi kullanarak ve MorphAnalyzer işlemi kullanmadan eğitilmiştir. MorphAnalyzer işlemi kullanarak ve kullanmadan olmak üzere her bir model için ayrı ayrı Accuracy, Precision, Recall ve F-1 Skorları hesaplanmıştır.

A. Sonuçların İncelenmesi

İlgili eğitim veri seti 6 farklı modelle eğitilmiş ve elde edilen sonuçlar aşağıdaki gibidir.

MorphAnalyzer Kullanarak				
	Accuracy	**Recall**	**Precision**	**F-1**
Model - 1	0.890	0.890	0.889	0.886
Model - 2	0.830	0.830	0.839	0.833
Model - 3	0.826	0.826	0.836	0.830
Model - 4	0.829	0.829	0.838	0.832
Model - 5	0.841	0.841	0.841	0.837
Model - 6	0.713	0.713	0.765	0.706

MorphAnalyzer Kullanmadan				
	Accuracy	**Recall**	**Precision**	**F-1**
Model - 1	0.911	0.911	0.918	0.900
Model - 2	0.813	0.813	0.825	0.814
Model - 3	0.807	0.807	0.821	0.809
Model - 4	0.808	0.808	0.821	0.809
Model - 5	0.821	0.821	0.829	0.817
Model - 6	0.724	0.724	0.776	0.716

İlgili sonuçlar incelendiğinde MorphAnalyzer işlemi yapılmadan eğitilen Model-1'in en yüksek doğruluk, kesinlik, duyarlılık ve F-1 skorunu verdiği görülmektedir. En iyiden en kötüye göre sıralama yapılacak ise Model-1, Model-5, Model-2, Model-3 ve Model-6 sıralaması yapılabilir. MorphAnalyzer işlemi kullanmadan eğitilen modellerde en yüksek sonuçları veren modelin yine Model-1 olduğu görülebilir. En iyiden en kötüye göre sıralama yapılmak istenirse de Model-1, Model-5, Model-2, Model-4, Model-3 ve Model-6 şeklinde sıralama yapılabilir.

B. Tartışma

Elde edilen sonuçlar ışığında Model-1'in en iyi sonucu verdiği bilinse de modellerin daha iyi sonuçlar verebilmesi adına ilgili veri seti üzerinde daha fazla çalışma yapılabilir. Örneğin kullanılan veri seti birçok elektronik mağazanın ürün yorumlarını içermektedir. Kullanıcılar tarafından yapılan yorumların yazım yanlışları ve ironi içerebileceği düşünüldüğünde daha az veri seti ile veya ilgili veri seti üzerinde yormların incelenip düzeltilmesi işlemlerinden sonra yeniden çalışılması ilgili modellerin daha iyi sonuçlar vermesini sağlayabilir. Bununla birlikte ilgili modeller farklı makine öğrenmesi algoritmaları ile, farklı max features değerleri ile ve farklı n-gram'larla yeniden test edilebilir

Kaynaklar

[1] Y. Bozdağ and O. Yalçınkaya Alkar, "Bergen alışveriş bağımlılığı Ölçeği'nin kompülsif Çevrimiçi satın alma davranışına uyarlanması," *Bağımlılık Dergisi*, vol. 19, no. 2, pp. 23–34, 2018.

[2] F. Bal and I. Okkay, "İnternet tabanlı sorunlu alışveriş davranışı: Çevrimiçi alışveriş bağımlılığı," *Bağımlılık Dergisi*, vol. 23, no. 1, pp. 111–120, 2022.
[3] G. Alparslan and M. Dursun, "Konvolüsyonel sinir ağları tabanlı türkçe metin sınıflandırma," *Bilişim Teknolojileri Dergisi*, vol. 16, no. 1, pp. 21 – 31, 2023.
[4] Y. Özkan and V. M. Yöntemleri, "Papatya yayıncılık eğitim," *İstanbul, Mayıs-2008*, 2013.
[5] M. Aydoğan and A. Karcı, "Meslek yüksekokulu öğrencilerinin başarı performanslarının makine öğrenmesi yöntemleri ile analizi," in *International Symposium on Multidisciplinary Studies and Innovative Technologies*, 2018.
[6] I. H. M. Paris, L. S. Affendey, and N. Mustapha, "Improving academic performance prediction using voting technique in data mining," *International Journal of Computer and Information Engineering*, vol. 4, no. 2, pp. 306–309, 2010.