All Sciences Proceedings
http://as-proceeding.com/

2nd International Conference on Engineering, Natural and Social Sciences

April 4-6, 2023 : Konya, Turkey

© 2023 Published by All Sciences Proceedings

https://www.icensos.com/

Yüksek Boyutlu Amazon Ürün Değerlendirmeleri Veri Kümesi Üzerinde Bayes Sınıflandırıcılarının Performans Karşılaştırması

Ensar Arif Sağbaş[1*]

[1]*Bilişim Sistemleri Mühendisliği Bölümü / Teknoloji Fakültesi, Muğla Sıtkı Koçman Üniversitesi, Türkiye*

[*]*arifsagbas@mu.edu.tr Başlıca yazarın mail adresi*

Özet – İnternet teknolojilerinin hızla gelişmesiyle birlikte elektronik belge sayısı dünya çapında büyük bir artış göstermiştir. Duygu analizi, çevrimiçi metin belgelerinden bilgilerin çıkarılması için kritik bir görevdir. Tipik olarak denetimli makine öğrenimi algoritmaları tarafından gerçekleştirilen duygu analizi, çevrimiçi metin belgelerinden öznel bilgileri çıkarmak için oldukça kullanışlıdır. Metin belgesi sınıflandırmasının temel amacı, elektronik metin belgelerine uygun sınıflar atamaktır. Sınıflandırmadaki kilit nokta ise veri kümesine uygun sınıflandırıcı yaklaşımına karar vermektir. Bu çalışmada, duygu sınıflandırmasında kullanılan, yüksek boyutlu Amazon ürün değerlendirmeleri veri kümeleri üzerinde Bayes lojistik regresyon, Bayes ağları, Naive Bayes ve çok terimli Naive Bayes olmak üzere dört farklı Bayes sınıflandırıcısının performansları test edilmiş ve karşılaştırılmıştır. Sınıflandırmalar veri kümesinin tamamına ek olarak, korelasyon tabanlı öznitelik seçimi ile oluşturulan beş yeni öznitelik alt kümesi ile gerçekleştirilmiştir. Her bir veri kümesi için dört Bayes sınıflandırıcı ve altı öznitelik alt kümesi kombinasyonu test edilmiştir. En başarılı sonuçlar bütün veri kümelerinde %90'ın üzerinde doğruluk oranı ile Bayes lojistik regresyon yöntemi ile elde edilmiştir.

Anahtar Kelimeler – Amazon Veri Kümesi, Bayes Sınıflandırıcıları, Çok Terimli Naive Bayes, Bayes Lojistik Regresyon, Duygu Analizi, Doğal Dil İşleme

I. GİRİŞ

İnternet teknolojilerinin hızlı yükselişiyle birlikte ürün incelemeleri, derecelendirmeler, tavsiyeler ve diğer kullanıcı tarafından oluşturulan kişisel görüşler, araştırma toplulukları tarafından büyük ilgi görmüştür [1]. Bu bilgilerin muazzam hacmi ve katlanarak büyümesi, hükümetlere, işletmelere ve kullanıcıların kendilerine potansiyel değerler sağlamaktadır [2]. Çevrimiçi olarak oluşturulan içeriğin büyük çoğunluğunda yer alan duygu adı verilen doğal bir özellik bulunmaktadır. Buradaki duygu, bir şey hakkında sahip olduğunuz bir fikir veya düşünce olarak tanımlanabilir [3]. Duyguların sınıflandırılması ise metin madenciliği alanında sıklıkla çalışılan ve başarı ile ele alınan bir konudur [4-6].

Makine öğrenimi, çevreleyen ortamdan öğrenerek insan zekâsını taklit etmek için tasarlanmış, gelişen bir hesaplama algoritmaları dalıdır. Makine öğrenimine dayalı teknikler, örüntü tanıma, bilgisayar görüşü, doğal dil işleme, uzay aracı mühendisliği, finans, eğlence ve hesaplamalı biyolojiden biyomedikal ve tıbbi uygulamalara kadar çeşitli alanlarda başarıyla uygulanmıştır [7]. Makine öğrenmesi üç kategoride ele alınabilir. Bunlar denetimli öğrenme, denetimsiz öğrenme ve pekiştirmeli öğrenmedir. Denetimli (supervised) öğrenme, basitçe örneklerden öğrenme fikrinin hayata geçirilmesidir [8]. Denetimli öğrenme algoritmaları, sınıflandırma ve regresyon gibi çeşitli görevler için kullanılabilir. İstatistikte sınıflandırma, bir gözlemin (veya gözlemlerin) bir dizi kategoriden (alt popülasyon) hangisine ait olduğunu belirleme görevidir [9]. Duyguların sınıflandırılmasında karar ağaçları [10,11] , Naive Bayes [4,12], destek vektör makinesi [11,12], lojistik regresyon [12] ve kNN [10-12] gibi bilinen denetimli makine öğrenmesi yöntemlerinden

faydalanılmaktadır. Bu çalışmada Blitzer vd. [13]'den sağlanan yüksek boyutlu metin duygu analizi veri kümesi üzerinde Bayes lojistik regresyon, Bayes ağları, Naive Bayes ve çok terimli Naive Bayes olmak üzere dört adet Bayes sınıflandırıcılarının performansları karşılaştırmalı olarak test edilmiş ve sonuçları değerlendirilmiştir.

II. MATERYAL VE YÖNTEM

A. *Veri Kümesi*

Bu çalışmada duyguların sınıflandırılması için Blitzer vd. [13]'den sağlanan 4 adet Amazon ürün incelemeleri veri kümesinden faydalanılmıştır. Bu veri kümelerinin isimleri books, dvd, electronics ve kitchen'dır ve %50 olumlu, %50 olumsuz gözlem içermektedir. Veri kümelerinin özellikleri Tablo 1'de, kısa açıklamaları Tablo 2'de gösterilmektedir.

Tablo 1. Kullanılan veri kümelerinin özellikleri

Veri kümesi	**Öznitelik sayısı**	**Gözlem sayısı**
books	28234	2000
dvd	28310	2000
electronics	14943	2000
kitchen	12130	2000

Tablo 2. Kullanılan veri kümelerinin açıklamaları

Veri kümesi	**Açıklama**
books	Amazon.com'dan kitap ürün incelemeleri
dvd	Amazon.com'dan DVD ürünleri incelemeleri
electronics	Amazon.com'dan elektronik ürünlerin incelemeleri
kitchen	Amazon.com'dan mutfak ürünleri incelemeleri

B. *Naive Bayes (NB)*

Naïve Bayes sınıflandırıcısı, sınıf bağlamında verilen tüm değişkenler arasında güçlü koşullu bağımsızlık varsayımı yapan olasılıklı ve basit bir sınıflandırıcıdır [14]. Belgelerin temel birimlerden nasıl oluştuğuna bağlı olarak, farklı Naïve Bayes olay modelleri bulunmaktadır [15].

Metin sınıflandırması için, bir d_j belgesinin c sınıfına ait olma olasılığı Denklem 1'deki Bayes teoremi tarafından hesaplanır.

$$p(c|d_j) = \frac{p(d_j|c)p(c)}{p(d_j)} \quad (1)$$

C. *Çok terimli Naive Bayes (ÇTNB)*

Çok terimli Naïve Bayes modeli, çok değişkenli Bernoulli olay modeli yerine belgelerin uzunluklarının belgelerdeki sınıftan bağımsız olduğunu varsayarak kelime frekans bilgisini kullanan üretken bir modeldir [14]. Sınıflandırıcı, kelime sıklık bilgisinden yararlandığı için metin sınıflandırma görevleri için çok uygundur.

D. *Bayes Lojistik Regresyon (BLR)*

Bayes lojistik regresyonu, lojistik regresyona Bayes yaklaşımı uygulayan bir sınıflandırıcıdır. Tahmin edici, aşırı öğrenmeyi önlemek için bir Laplace prior kullanır ve metin verileri için seyrek sınıflandırma modelleri oluşturur. Bu model en az öznitelik seçimi ile birleştirilmiş destek vektör makinesi veya ridge lojistik regresyon kadar güçlüdür [16].

E. *Bayes Ağları (BA)*

Bayes ağları, olasılıksal çizge modellerden biridir. Bayes ağlarında belirsiz bir konu hakkındaki bilgi, çizgesel yapılar olarak gösterilir. Özellikle, değişkenler çizgede düğümler olarak temsil edilirken, değişkenler arasındaki olasılıksal bağımlılıklar kenarlar olarak temsil edilir. Çizgedeki kenarların değerleri, bilinen hesaplama ve istatistiksel yöntemler kullanılarak hesaplanabilir [17,18].

F. *Korelasyon Tabanlı Öznitelik Seçimi (KTÖS)*

Korelasyon ölçüsü, özellik ile sınıf arasındaki Pearson korelasyon katsayısını ölçerek bir özelliği değerlendirir. Bu ölçüm, özellik ile sınıfın arasındaki korelasyonun gücünü temsil eder. Korelasyon katsayısı +1 ile -1 arasında bir değere sahiptir. +1 toplam pozitif lineer korelasyonu gösterir, 0 lineer korelasyon olmadığını gösterir ve -1 toplam negatif lineer korelasyonu gösterir. Denklem 2 de gösterilen formül ile hesaplanır [19].

$$R(i) = \frac{\sum_{k=1}^{m}(x_{k,i} - \bar{x}_i)(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(x_{k,i} - \bar{x}_i)^2 . \sum_{k=1}^{m}(y_k - \bar{y})^2}} \quad (2)$$

m veri noktalarının sayısı, x öznitelik ve y sınıftır.

III. BULGULAR

Yüksek boyutlu amazon ürün incelemeleri veri kümesi üzerinde dört farklı Bayes sınıflandırıcının (Bayes lojistik regresyonu, Bayes ağları, Naive Bayes ve çok terimli Naive Bayes) performansı test edilmiştir. Her bir veri kümesinin öznitelikleri korelasyon tabanlı öznitelik seçimi algoritması ile derecelendirilmiştir. En iyi puana sahip 300, 750, 1500, 3000 ve 6000 adet öznitelik birleştirilerek yeni öznitelik alt kümeleri elde edilmiştir. Sonrasında elde edilen yeni veri kümeleri Bayes sınıflandırıcıları ile sınıflandırılmış ve performansları karşılaştırılmıştır. Deneyler WEKA [20] yazılım aracı ile on katmanlı çapraz doğrulama uygulanarak gerçekleştirilmiştir. Sınıflandırıcıların parametre değerleri olarak WEKA aracının varsayılan değerleri kullanılmıştır. Books veri kümesi için elde edilen sınıflandırma doğruluk oranları Tablo 3'te sunulmuştur.

Tablo 3. Books veri kümesi için elde edilen sonuçlar

#Öznitelik	BLR	BA	NB	ÇTNB
300	86.05	**74.65**	**74.60**	85.20
750	87.95	74.40	74.45	86.85
1500	90.20	74.40	74.45	87.80
3000	**90.60**	74.40	74.45	89.05
6000	90.30	74.40	74.45	**89.80**
Tümü	79.30	74.40	74.45	81.75

Books veri kümesinde en yüksek sınıflandırma başarısı BLR yöntemi ile elde edilmiştir. Bu sınıflandırma için 3000 öznitelikten oluşan öznitelik alt kümesi kullanılmıştır. Sonrasında bu yöntemi %89.80 ile ÇTNB takip etmiştir. BA ve NB yöntemleri ise bu iki yönteme göre rekabetçi sonuçlar üretememiştir. Dvd veri kümesi için elde edilen sınıflandırma doğruluk oranları Tablo 4'te verilmiştir.

Tablo 4. DVD veri kümesi için elde edilen sonuçlar

#Öznitelik	BLR	BA	NB	ÇTNB
300	87.20	78.20	78.20	86.35
750	89.25	78.30	78.20	87.60
1500	90.45	**78.45**	**78.55**	88.80
3000	90.60	78.45	78.55	**89.75**
6000	**90.80**	78.30	78.45	89.50
Tümü	82.50	78.30	78.45	83.35

Dvd veri kümesinde de books veri kümesine benzer sonuçlar elde edilmiştir. En yüksek sınıflandırma başarısını sağlayan yöntemler sırasıyla BLR ve ÇTNB olmuştur. BA en yüksek %78.45 doğruluk oranına ulaşırken NB %78.55 doğruluk oranını yakalamıştır. Electronics veri kümesi için elde edilen sınıflandırma doğruluk oranları Tablo 5'te sunulmuştur.

Tablo 5. Electronics veri kümesi için elde edilen sonuçlar

#Öznitelik	BLR	BA	NB	ÇTNB
300	88.20	**80.90**	**81.05**	86.55
750	88.95	80.80	80.90	87.80
1500	**90.00**	80.75	80.85	**88.45**
3000	88.50	80.75	80.85	87.80
6000	88.30	80.75	80.85	88.25
Tümü	83.65	80.75	80.85	83.00

Electronics veri kümesinde de en başarılı sonuçlar BLR yöntemi ile elde edilmiştir. 1500 adet öznitelikte ile oluşturulan alt kümenin sınıflandırılması sonucu %90 başarı sağlanmıştır. Aynı öznitelik alt kümesinin ÇTNB yöntemi ile sınıflandırılması sonucunda ise %88.45 doğruluk oranı yakalanmıştır. Kitchen veri kümesi için elde edilen sınıflandırma doğruluk oranları Tablo 6'da verilmiştir.

Tablo 6. Kitchen veri kümesi için elde edilen sonuçlar

#Öznitelik	BLR	BA	NB	ÇTNB
300	90.25	**83.55**	83.45	89.10
750	91.30	83.55	**83.50**	**90.30**
1500	91.35	83.55	83.50	89.60
3000	**92.15**	83.55	83.50	89.30
6000	91.95	83.55	83.50	89.30
Tümü	88.35	83.55	83.50	85.60

Kitchen veri kümesinde ise en yüksek sınıflandırma oranı 3000 elemanlı öznitelik alt kümesi ile BLR yöntemi kullanılarak sağlanmıştır. ÇTNB sınıflandırıcısı ise 750 adet öznitelik ile oluşturulan alt küme ile en yüksek sınıflandırma başarısına ulaşmıştır. Diğer veri kümelerinde olduğu gibi BA ve NB yöntemleri bu veri kümesinde de ÇTNB ve BLR yöntemleri ile rekabet edememiştir.

IV. TARTIŞMA VE SONUÇLAR

Bu çalışmada duygu sınıflandırılmasında kullanılan, dört adet yüksek boyutlu Amazon ürün incelemeleri veri kümeleri üzerinde dört farklı Bayes sınıflandırıcı yaklaşımının performansları değerlendirilmiştir. Veri kümelerindeki öznitelikler korelasyon tabanlı öznitelik seçimi yöntemi ile derecelendirilmiş ve elde edilen puanlara göre en iyi 300, 750, 1500, 3000 ve 6000 adet öznitelik birleştirilerek yeni öznitelik alt kümeleri elde edilmiştir. Her bir veri kümesi, oluşturulan bu yeni 5 adet öznitelik alt kümesi ve ek olarak veri kümesinin tamamı ile dört Bayes sınıflandırıcı kullanılarak sınıflandırılmıştır. Elde edilen deneysel bulgular sonucunda dört veri kümesinde de en başarılı sınıflandırma sonuçları BLR yönteminden elde edilmiştir. Books ve kitchen veri kümeleri için 3000, electronics veri kümesi için 1500 ve dvd veri kümesi için 6000 elemanlı öznitelik alt kümesi kullanılmıştır. Bu yöntemi dört veri kümesinde de ÇTNB yöntemi takip etmiştir. Deneylerde göreceli olarak ÇTNB sınıflandırıcısı daha hızlı çalışıyor olsa da Amazon ürün incelemeleri veri kümesinde BLR yöntemi kadar başarılı sonuçlar sağlayamamıştır. BA ve NB yöntemleri ise hiçbir veri kümesinde ÇTNB ve BLR yöntemlerine rekabet edememiştir. BA yöntemi varsayılan parametre değerleri kullanıldığında NB yöntemine benzemektedir. Deneyler sonucunda da elde edilen paralel ve çok yakın sınıflandırma sonuçları bunu desteklemektedir. NB ve BA yöntemlerinden daha başarılı sonuçların elde edilebilmesi için daha ayrıntılı ve kapsamlı bir öznitelik seçimi yaklaşımına ihtiyaç duyulduğu düşünülmektedir. Yöntemlerin teorik anlatımında da ele alındığı gibi ÇTNB yöntemi metin sınıflandırma görevlerindeki başarısını göstermiştir. Aynı şekilde BLR yöntemi ise öznitelik seçimi uygulanmış bir destek vektör makinesi gibi çalışabileceğine değinilmiştir. Bu yöntem ele alınan diğer Bayes yaklaşımları arasında en başarılısı olmuştur. Gelecek çalışmalarda, BLR yönteminin sarmalayıcı tabanlı daha gelişmiş bir öznitelik seçimi yaklaşımı ile birleştirilerek çok daha başarılı sonuçların elde edilmesi planlanmaktadır.

KAYNAKLAR

[1] B. Pang, and L. Lee, "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval*, vol. 2, pp. 1–135, 2008.

[2] Y. He, and D. Zhou, "Self-training from labeled features for sentiment analysis", *Information Processing & Management*, vol. 47, no. 4, pp. 606-616, 2011.

[3] L. Delacroix, *Longman advanced American dictionary*, Edinburgh, UK: Pearson Education, 2007.

[4] O. Gokalp, E. Tasci, and A. Ugur, "A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification", *Expert Systems with Applications*, vol. 146, no. 113176, 2020.

[5] A. Onan, S. Korukoğlu, and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification", *Information Processing & Management*, vol. 53, no. 4, pp. 814-833, 2017.

[6] E. A. Sağbaş, "Filtre Tabanlı Öznitelik Seçim Yöntemleri Kullanılarak Metinlerde Duygu Sınıflandırması Üzerine Karşılaştırmalı Bir Çalışma", *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 35, no. 1, pp. 239-250, 2023.

[7] I. El Naqa, and M. J. Murphy, *What is machine learning?*, Springer International Publishing, pp. 3-11, 2015.

[8] E. G. Learned-Miller, "Introduction to supervised learning", *I: Department of Computer Science, University of Massachusetts*, 3, 2014.

[9] (2023) Statistical classification. [Online]. Available: https://www.wikiwand.com/en/Statistical_classification

[10] G. Wang, Z. Zhang, J. Sun, S. Yang, and C. A. Larson, "POS-RS: A Random Subspace method for sentiment classification based on part-of-speech analysis", *Information Processing & Management*, vol. 51, no. 4, pp. 458-479, 2015.

[11] A. Onan, S. Korukoğlu, and H. Bulut, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification", *Expert Systems with Applications*, vol. 62, pp. 1-16, 2016.

[12] A. Jalilvand, and N. Salim, "Feature unionization: a novel approach for dimension reduction", *Applied Soft Computing*, vol. 52, pp. 1253-1261, 2017.

[13] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification", In *Proceedings of the 45th annual meeting of the association of computational linguistics*, 2007, pp. 440-447.

[14] A. McCallum, and K. Nigam, "A comparison of event models for naive bayes text classification", In *AAAI-98 workshop on learning for text categorization*, 1998, vol. 752, no. 1, pp. 41-48.

[15] K. M. Schneider, "Techniques for improving the performance of naive bayes for text classification", In *Computational Linguistics and Intelligent Text Processing: 6th International Conference*, CICLing 2005, pp. 682-693.

[16] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale Bayesian logistic regression for text categorization", *technometrics*, vol. 49, no. 3, pp. 291-304, 2007.

[17] I. Ben-Gal, "Bayesian Networks", In: *Fabrizio Ruggeri, Ron Kenett and Frederick Faltin, editors. Encyclopedia of Statistics in Quality and Reliability*, Chichester, UK: Wiley; 2007.

[18] S. Balli, and E. A. Sağbaş, “The usage of statistical learning methods on wearable devices and a case study: activity recognition on smartwatches”, *Advances in statistical methodologies and their application to real problems*, pp. 259-277, 2017.
[19] X. W. Chen, and M. Wasikowski, “Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems”, In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 124-132.
[20] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, Morgan Kaufmann, Fourth Edition, 2016.