Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 4(1) (2021) 8-16
DOI: 10.38016/jista.801300

Tüketici Yorumları Üzerine Bir Metin Madenciliği ve Veri Boyutu İndirgeme Yaklaşımı

Ahmet YÜCEL*

Ankara Yıldırım Beyazıt Üniversitesi, Ankara, Türkiye
ayucel@ybu.edu.tr

Öz

Veri boyutunun artmasıyla orantılı olarak değişkenler arası bağlantıların analizi daha karmaşık bir hale gelmiştir. Yapısal olmayan veri kümelerinde, yapısal forma dönüştürme ön işlemleriyle birlikte, analiz süreçleri daha karmaşık hale gelecektir. Konuşma dilinin doğası gereği, sıradan bir doküman dahi yüzlerce farklı terim içermektedir. Bu durum veri çıkarımı ve verinin yapısal forma dönüştürme süreçlerini oldukça uzatmaktadır. Bu çalışmada kullanılan veri, E-ticaret olarak adlandırılan, çevrimiçi alışveriş işlemleri sırasında ve sonrasında, gerçek kişiler tarafından yazılan yorumlardan oluşmaktadır. Alternatif bir alışveriş yöntemi olan e-ticaret platformlarında, tüketiciler istedikleri ürüne/hizmete ait birçok farklı seçeneği eşzamanlı inceleyebilmektedir. Tüketiciler bu sayede almış oldukları ürünle/hizmetle alakalı tecrübelerini/düşüncelerini kolayca ifade edebilirken, diğer tüketicilerin yorumlarına ulaşma fırsatını da bulabilmektedir. Bu durum metin veri açısından, sürekli büyüyen bir kaynak oluşturmaktadır. Veri boyutunun sürekli olarak artması, veri analizindeki zorluğu da aynı oranda arttırmaktadır. Boyut problemini aşmak için metin madenciliği (MM) çalışmalarında oldukça popüler olan veri boyutu indirgeme yöntemlerinden biri, Tekil Değer Ayrışımı (TDA) kullanılmaktadır. Bu çalışmada, sınıflandırmaya dayalı polarite yöntemi, kompozit (bileşik) bir değişken oluşturma sürecinde kullanılmaktadır. Oluşturulan kompozit değişken, veri içinde yer alan tüm kelime ve kelime gruplarının matematiksel olarak bir araya getirilmesiyle oluşmaktadır. Dolayısıyla ilgili değişken bir veri boyutu indirgeme fonksiyonu da sağlamaktadır. TDA ve kompozit değişkenin, veri boyutu indirgeme performansları kıyaslanmaktadır. Modelleme yöntemi olarak, Genelleştirilmiş Lineer Model (GLM) kullanılmaktadır. Modellerin performansları 5-katmanlı-çapraz-doğrulama yöntemiyle değerlendirilmektedir. TDA skorları ve kompozit değişken kullanılarak GLM modelleri oluşturulmaktadır. Sonuçlar, 5-katmanın tamamında kompozit değişkenin TDA skorlarından ortalama %6 civarında daha iyi performans sağladığını göstermektedir. Bu yaklaşım, MM'nin veri analizi sürecini kolaylaştırmada ve doğruluk performansını arttırmada önemli bir katkı sağlayacaktır.

Anahtar kelimeler: Metin madenciliği, Polarite, Tekil Değer Ayrışımı.

Text Mining and Data Dimesion Reduction Approach on Consumer Comments

Abstract

In proportion to the increase in data size, the analysis of connections between variables has become more complex. In unstructured datasets, analysis processes would become even more complex with transformation preprocessing to structural form. Due to the nature of natural speaking language, even an ordinary document contains hundreds of different terms. This situation extends the processes of data extraction and transformation. The data used in this study consists of comments written by real persons during and after online shopping, called e-commerce. On e-commerce platforms, which are an alternative shopping method, consumers can simultaneously examine many different options of a product / servic. In this way, consumers can easily express their experiences / thoughts about the product / service they have purchased, and have the opportunity to access other consumers' comments. This situation creates an continously-growing source of text data. The continuous increasing in data size increases the difficulty in data analysis at the same rate. One of the most popular data size reduction methods in text mining (MM) studies, Singular Value Decomposition (SVD) is used to overcome the size problem. In this study, the classification-based polarity method is used in the process of creating a composite variable. The composite variable is formed by mathematically combining all words and phrases in the document. Therefore, the relevant variable also provides a data size reduction function. Data size reduction performances of the SVD and the composite variable are compared. Generalized Linear Model (GLM) is used as the modeling method. The performances of the models are evaluated with 5-fold-cross-validation method. GLM models are created by using TDA scores and composite variables. The results show that the

* Sorumlu yazar: Ahmet YÜCEL
E-posta adresi: ayucel@ybu.edu.tr

Alındı : 28 Eylül 2020
Revizyon : 24 Kasım 2020
Kabul : 26 Kasım 2020

composite variable outperformed the SVD scores on average by about 6% in the all 5-layers. This approach will make a significant contribution in facilitating the data analysis process of MM and increasing its accuracy performance.
Keywords: Text Mining, Polarity, Singular Value Decomposition.

1. Giriş (Introduction)

Bilgisayar ve iletişim teknolojileri büyük bir hızla gelişimini sürdürürken, sosyal hayatın her aşamasında da aynı hızla yerini almaktadır. Günlük hayatın doğal akışı, gittikçe daha da artan bir şekilde elektronik bir form kazanmaktadır. Ekonomiden sağlığa, hemen her alanda devam eden süreçler, anlık olarak bilgisayar ortamında kayıt altına alınmaktadır. Bu durum, anlık kayıtlı veri oluşumunun büyüklüğünü ve hızını insan hayal gücünün de ötesine taşırken, veri madenciliği ve veri analizi yapmak için mükemmel bir imkân sağlamaktadır. Sürekli olarak artan veri boyutları, veri analizi süreçlerini oldukça güçleştirmektedir. Bu amaçla mevcut yöntemlere ek olarak birçok yeni veri boyutu indirgeme yöntemi geliştirilmektedir (Varghese, 2012).

Bir verinin kalitesini belirleyen unsurların başında verinin ulaşılabilirliği, uygun büyüklüğe sahip oluşu, gerçekliği, ilgili alanı kapsayıcılığı, objektifliği ve güncelliği gelmektedir (Pipino, 2002). Bu unsurlar dikkate alındığında, sosyal platformlarda oluşan verinin kıymeti ve sağlayabileceği bilginin kalitesi açıkça anlaşılmaktadır. Elektronik ortamda oluşan verinin %80'den fazlasının yapısal olmayan formda olduğu bilinmektedir (Rajalakshmi, 2015). Bu durum, yapısal olmayan bir veriyi yapısal forma dönüştürmede kullanılan metin madenciliğinin de önemini ortaya koymaktadır. Bu çalışmamızda, e-ticaret alanında, tüketici yorumlarına dayalı bir metin madenciliği uygulamasını göstereceğiz. Bu sayede, TDA'nın ve polariteye dayalı geliştirilen bir kompozit değişkenin veri boyutu indirgemedeki performanslarını inceleyeceğiz.

Elektronik ticaret (E-ticaret) elektronik iletişim yollarıyla, ürün ve hizmetlerin verilmesi anlamına gelmektedir. Her ne kadar 2000'li yıllardan sonra yaygınlığı artsa da, e-ticaretin ortaya çıkışı 30 yıl geriye kadar uzanıyor. Elektronik veri değişimi (EVD), elektronik ortamda belgelerinin bir bilgisayardan diğerine standart bir formatta ulaştırılabilmesi yöntemi, e-ticaretin yapılabilmesini mümkün kılan temel başlangıç noktası olmuştur. EVD, 1960'lı yıllarda, bazı taşımacılık ve perakendecilik sektörlerinin kırtasiye masraflarını azaltma çabasıyla ortaya çıkmıştır. 1990'lara gelindiğinde, halen Avrupa'daki ve Amerika Birleşik Devletleri'ndeki şirketlerin çok azı, EVD yöntemini kullanmaya başlamıştı. Ancak bu daha sonra geliştirilen World Wide Web (WWW) ve bununla ilişkili olarak kaynak adres bilgisi veren URL ve HTML dili e-ticaretin bugünkü seviyesine ulaşmasının önünü açmıştır (Tian, 2008). E-ticaretle birlikte kişi ve işletmelerin alış/satış, tedarikçileri ve müşterileri ile ilişkilerini yönetme, lojistik ve envanter düzenleme süreçleri hızlanmış ve kolaylaşmıştır. Dolayısıyla maliyetler de önemli ölçüde azalmıştır. Tüm bu sağladığı kolaylık ve hızla, e-ticaret yeni bir rekabet alanı oluşturmuştur. Bu amaçla Joyo Amazon, DangDang gibi birçok şirket e-ticaret sektörüne güçlerini birleştirerek girmiştir. Tekstil endüstrisi de geleneksel pazarlama yöntemini, e-ticaretin ortaya çıkması nedeniyle değiştirmiştir. Zhenxiang ve Lijie çalışmalarında Çinli Zara ve Vancl tekstil firmalarının online perakendecilik alanında bir araya gelmelerini ve bu alanda gösterdikleri başarılarını incelemişlerdir (Zhenxiang, 2011).

Uzun süre sınırlı alanlarda kullanılan e-ticaret, bugün bütün sektörlerde kullanılmaya başlanmıştır. Bu durum müşteri etkileşimlerini kayıt altına alma imkânı sağlamıştır. Bu amaçla hizmet etkileşimlerinde müşterilerin yorum ve görüşlerini incelemek için duygu analizi teknikleri geliştirilmiştir. 2018'de Yom-Tov tarafından yapılan bir çalışmada müşterilerin duygu analizini otomatik olarak yapacak bir algoritma geliştirilmiştir. Bu sayede, müşteri görüş ve duyguları dinamik bir biçimde takip edilmiştir. Başlangıçta negatif müşteri yorumlarına konu olan aksaklıklar belirlenmiş ve sonrasında ilgili firmaların bu alanlar üzerine yaptıkları düzeltmelerle, müşteri yorumlarının pozitife döndüğü gözlemlenmiştir. Bu sayede, hizmet etkileşimi sırasında müşteri duygu dinamikleri ile hizmet başarısızlığı ve iyileştirme kavramları arasındaki ilişki unsurları belirlenmiştir. Bu durum gelecekteki hizmet kalitesini arttırmaya yardımcı bir bilgi kaynağı olarak değerlendirilmiştir. Yom-Tov çalışmasında, müşteri etkileşimi ile hizmet kalitesi arasındaki bağlantıyı net bir biçimde ortaya koymuştur ve web tabanlı hizmet kalitesinin gerçek zamanlı izlenebilmesi ve kontrolü için duyarlılık analiz araçlarının kullanılmasını önermiştir (Yom-Tov, 2018). Elbette e-ticaret alanında faaliyet gösteren firmalar müşterilerin duygu ve eğilimlerini takip etmek için geleneksel yöntemlerin yanı sıra, sosyal medyayı da aktif olarak kullanmaktadır. Bu amaçla Twitter gibi sosyal medya platformları önemli bir bilgi kaynağı oluşturmaktadır. Firmalar kendilerini ilgilendiren bir konuda mevcut olan bir başlık etiketi (hashtag) altına yapılan paylaşımları takip ederek veya kendilerinin belirleyeceği bir başlık etiketi yardımıyla, insanların firma tarafından paylaşılan bazı sorulara cevaplarını veya genel yorumlarını alması, firmanın müşterileriyle dinamik ve eş zamanlı bir etkileşim kurmasını sağlayacaktır. Buna ek olarak müşteriler tarafından paylaşılan bu veri, diğer tüketiciler için hizmet ve ürünler hakkında fikir sahibi olmalarını sağlayacaktır. Yani, bu değerli veriler hem firmanın hem de tüketicinin kararlarını desteklemek için kullanılabilmektedir. Al-Otaibi konuyla alakalı çalışmasında, metin formatında Twitter verilerini, destek vektör makinesi (Support Vector Machine

(SVM)) algoritması ile pozitif veya negatif şeklinde sınıflandırmıştır (Al-Otaibi, 2018). İnternet kullanımının yaygınlaşması ve internet ortamında kişisel fikir ifade etme kolaylığı, internet üzerinden algı çalışması yapabilmek için çok verimli bir alan olmuştur. Son on yılda söz konusu metin verilerin otomatik şekilde toplanması ve işlenmesi üzerine birçok duygu analizi algoritması geliştirilmiştir. Bu algoritmaların temel amacı bir metnin öznel olup olmadığını belirlemek ve eğer öznel ise, olumlu veya olumsuz bir görüş olduğunu tespit etmektir (Pajupuu, 2016).

Duygu analizinin temel amacı bireylerin bir konu hakkındaki eğilimlerini tespit etmek ve gelecek eğilimleri hakkında öngörü geliştirmektir. Pajupuu yaptığı çalışmada öznel metinlerin polaritesini otomatik olarak belirleyen makine öğrenimine ve kelime karakter dizilimine (lexical sequence) dayalı algoritmalar geliştirmiştir ve iki modelin performanslarını kıyaslamıştır. İlgili çalışmada makine öğrenimi yöntemi daha başarılı sonuç üretmiştir (Pajupuu, 2016). Benzer bir polarite çalışması da Singh tarafından yapılmıştır. Yapılan çalışmada Twitter'dan blog paylaşımlarına birçok farklı kaynaktan veri toplanmış ve duygu analizinde makine öğrenimine ve kelime karakter dizilimine (lexical sequence) dayalı algoritmaların performansları incelenmiştir. Ayrıca yapılan çalışmada gözetimli (supervised) ve gözetimsiz (unsupervised) algoritmalar da karşılaştırılmıştır. Çeşitli kaynaklardan elde edilen metinlerin duygu analizi ve sezgisel tabanlı duyarlılık modellemesi için makine öğrenimi algoritmalarına ek olarak, sözlüğe dayalı yöntemlerin performansı da değerlendirmektedir. Bu sayede sinema filmi veya benzeri ürünler için yapılan yorumların polarizasyonu sağlanmıştır. Her iki yaklaşımın kıyaslaması da ek olarak yapılmıştır (Singh, 2014). Singh'in sözlüğe dayalı yöntemlerin polarizasyon performansını makina öğrenimi algoritmalarının performanslarıyla kıyaslamıştır. Bunun yanında, makina öğrenimi algoritmalarının da kendi içinde performans olarak ayrıştığını tespit etmiştir. Zubrinic, beş farklı makine öğrenimi algoritmasını duygu analizi modellerinde uygulamıştır. Çevrimiçi müşteri yorumlarının polarizasyonuna dayalı bu çalışmada, Naive Bayes, Destek Vektör Makinesi, Yapay Sinir Ağları ve Maksimum Entropi C4.5 algoritmalarının performansı gözlemlenmiştir. Sonuçlar Destek Vektör Makinesi ve Maksimum Entropi C4.5 yöntemlerinin daha yüksek doğruluğa ulaştığını göstermektedir (Zubrinic, 2018). Zhao ve Xu ise çalışmalarında Yapay Sinir Ağlarının özel bir yaklaşımını doğal dil işleme sürecinde uygulamıştır. Çalışmada duygusal polarite ve yapay sinir ağları bir metin içinde yer alan çeldirici/manipüle edici ifadelerin tespitinde kullanılmıştır (Zhao, 2018).

Metin madenciliğinin en temel amacı, bilgi kaybetmeden metin içinde yer alan sınıfsal ve anlamsal bağlantıları belirlemektir. Ancak veri boyutunun hızla artması bu konuda önemli bir zorluk oluşturmaktadır. Arunachalam yaptığı çalışmada Bayesian sınıflandırmadan genetik sınıflandırmaya birçok sınıflandırma teknikleri üzerine fikirler geliştirmiştir (Arunachalam, 2017). Boling ve Das yaptıkları metin sınıflaması çalışmasında veri boyutunun oluşturduğu zorlukları aşmak için Tekil Değer Ayrışımı (TDA) veri indirgeme yöntemini kullanmıştır. TDA yönteminin kullanılmasının en temel sebebi, büyük boyutlu bir verinin getirdiği veri önişleme sürecinin zorluklarını azaltmak ve bunu yaparken orijinal veriden bilgi kaybını en az seviyede tutmaktır (Boling, 2015).

Bu çalışmada çevrimiçi alışveriş işlemleri sırasında ve sonrasında, paylaşılan müşteri yorumları üzerine duygu analizi yapılmıştır. Yöntem olarak sınıflandırmaya dayalı polarite ve kompozit değişkene dayalı boyut indirgeme yaklaşımları uygulanmıştır. Genelleştirilmiş Lineer Modeller (GLM) üzerinden, kompozit değişkenin sınıflandırma performansı ve veri boyutu indirgeme performansı TDA ile kıyaslanmıştır. Kompozit değişkenin TDA skorlarından ortalama %6 civarında daha iyi performans sağladığı gözlemlenmiştir. Bu yaklaşımla, metinsel veri analizi sürecinin kolaylaştırılması ve doğruluk performansının arttırılması hedeflenmiştir.

2. Materyal ve Yöntem **(Material and Method)**

2.1. Polarite (Polarity)

Polarite kavramı, çalışmaların içeriğine bağlı olarak farklı amaçla kullanıma sahip olsa da, bu çalışmada, bir dokümanı oluşturan öznel terimlerin olumlu ya da olumsuz olarak sınıflandırılması anlamında kullanılmıştır. Tomar yaptığı çalışmada polariteyi iki aşamalı olarak ele almıştır. Bunlar, bir terimin ön polaritesi ve bağlamsal polaritesi olarak ifade edilebilir. Ön polarite, terimin yalın halde taşıdığı anlama göre, bağıl polarite ise terimin bulunduğu cümle içinde taşıdığı anlama göre ele alınması olarak özetlenebilir (Tomar, 2016). Bağıl polarite sürecine analizcinin sezgisel olarak katlımı gerekmektedir. Bu da tamamen otomatik bir süreci imkânsız kılmaktadır. Bu sebeple çalışmada ön polariteye dayalı bir metin işleme süreci uygulanmıştır.

Çalışmanın bu kısmı, kişilerin duyguları ile terimler arasında ön polariteye dayalı bir ilişki tespiti hakkındadır. Temel amaç, her bir dokümanın içerdiği pozitif ("tavsiye edilen" recommended) ve negatif ("tavsiye edilmeyen" not recommended) terimlerin tespit edilmesi ve daha sonra mevcut bilgiyi matematiksel olarak dönüştürüp ilgili dokümanın genel pozitif oranının hesaplanmasıdır. Bu hesabın temel amacı ise, verinin toplandığı alanla ilgili polariteye dayalı bir sözlük oluşturmak ve her bir terimin taşıdığı polarite ağırlığının büyüklüğünü ve yönünü belirlemektir. Böylece ilgili alanda yer alan benzer verilerin otomatik duygu analizi yapılabilir olacaktır. Bu amaçla, her bir dokümanın pozitif polarite oranını içeren

kompozit (bileşik) bir değişken hesaplanmıştır. Polariteye dayalı oluşturulan kompozit değişkeninin matematiksel ifadesi aşağıdaki verilmiştir: Kullanılan veride mevcut bağımlı ve bağımsız değişkenler ikili (binary) formdadır. Yani,

$$a_{ij} = \begin{cases} 1, & \text{j. terim i. dokümanda mevcut} \\ 0, & \text{j. terim i. dokümanda mevcut değil} \end{cases} \quad (1)$$

Ayrıca, modelin başarısını en iyi şekilde test edebilmek için, bağımlı değişkenin her iki kategorisi için de eşit sayıda durum alınmıştır. Bu sebeple, yapısal olmayan veriden elde edilen terim sayısı n olmak üzere, oluşturulan yapısal verinin matris boyutu $(2k) \times n$ şeklinde ifade edilmiştir.

$$\boldsymbol{X_{(2k)\times n}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kn} \\ a_{(k+1)1} & a_{(k+1)2} & \cdots & a_{(k+1)n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{(2k)1} & a_{(2k)2} & \cdots & a_{(2k)n} \end{bmatrix} \quad (2)$$

$\boldsymbol{X_{(2k)\times n}}$ metin veriden elde edilen yapısal veri olsun öyle ki, ilk k adet satır bağımlı değişkenin '0' ("Not Recommended") kategorisine, ikinci k adet satır ise bağımlı değişkenin '1' ("Recommended") kategorisine karşılık gelsin. Yani,

$$\boldsymbol{A_{k\times n}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kn} \end{bmatrix} \quad (3)$$

$$\boldsymbol{B_{(2k)\times n}} = \begin{bmatrix} a_{(k+1)1} & a_{(k+1)2} & \cdots & a_{(k+1)n} \\ a_{(k+2)1} & a_{(k+2)2} & \cdots & a_{(k+2)n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{(2k)1} & a_{(2k)2} & \cdots & a_{(2k)n} \end{bmatrix} \quad (4)$$

olmak üzere, $\boldsymbol{A_{k\times n}}$ matrisi bağımlı değişkenin '0' kategorisine, $\boldsymbol{B_{(2k)\times n}}$ matrisi ise bağımlı değişkenin '1' kategorisine karşılık gelen kısım olsun. Ayrıca V_j, $\boldsymbol{X_{(2k)\times n}}$ matrisinin j. sütunudur öyle ki, yapısal olmayan veriden elde edilmiş j. değişkene (terime) karşılık gelmektedir ve değişkene ait birim değerler sırasıyla $\{a_{1j}, a_{2j}, \ldots, a_{kj}\}$ şeklinde verilmektedir.

$$\mathbf{0}_j = \sum_{i=1}^{k} a_{ij} \quad (5)$$

Öyle ki $\mathbf{0}_j$ j. terimin yer aldığı '0' kategorili doküman sayısıdır. Benzer olarak,

$$\mathbf{1}_j = \sum_{i=k+1}^{2k} a_{ij} \quad (6)$$

Öyle ki $\mathbf{1}_j$ j. terimin yer aldığı '1' kategorili doküman sayısıdır. Buna göre, j. terimin tüm veri için, '1' kategorili dokümanlar içinde yer alma oranı $\mathbf{1R}_j$ olarak hesaplanır.

$$\mathbf{1R}_j = \frac{\mathbf{1}_j}{\mathbf{0}_j + \mathbf{1}_j} \quad (7)$$

Her bir dokümanın $(i = 1,2,\ldots,2k)$ bireysel 'kategori 1' eğilimini tespit etmek için $Oran1\_i$ değerini hesaplanacaktır. Bu amaçla öncelikle her dokümandan elde edilen toplam terim sayısı $(\boldsymbol{TS}_j)$ hesaplanmıştır.

$$\boldsymbol{TS}_j = \sum_{j=1}^{n} a_{ij} \quad (8)$$

$$\boldsymbol{Oran1\_i} = \frac{\sum_{j=1}^{n} (\mathbf{1R}_j \cdot a_{ij})}{\boldsymbol{TS}_j} \quad (9)$$

Polariteye dayalı değişken Review_Rec1_Rate_N'e (N Fold (katman) Sayısı (k-fold cross-validation)) ait birim değerleri *Oran1_1, Oran1_2, …, Oran1_2k* şeklinde ifade edilmektedir.

2.2. Tekil değer ayrışımı (TDA)

Tekil Değer Ayrışımı (TDA) çok popüler bir matris boyutu küçültme tekniğidir ve istatistikte büyük boyutlu veri kümelerinin boyutlarının bilgi kaybı oluşmadan, daha makul seviyelere indirgenmesi için kullanılmaktadır. Metin madenciliğinin doğası gereği, küçük sayılabilecek metinlerden dahi binlerce öznel terim çıkarımı olabildiğinden, TDA metin madenciliği alanında çok yaygın kullanıma sahip bir yöntemdir. TDA indirgeme işlemi özetle şu şekilde yapılmaktadır: A matrisi m×n boyutunda, metin veriden üretilmiş ikili (binary) formda bir (terim-doküman) frekans dağılım tablosu olsun öyle ki m doküman (tüketici yorum) sayısı ve n çıkarılan (seçilmiş) terimlerin sayısıdır.

$$A = UDV' \quad (10)$$

Öyle ki *U*, $m \times r$ boyutunda ortogonal bir matris, V, $n \times r$ boyutunda ortogonal bir matris, *V'* matrisi *V* matrisinin eşlenik transpozesi ve *D* $r \times r$ boyutunda kare bir matristir. Burada r *A'A* çarpımının özdeğer (eigenvalue) sayısıdır öyle ki *A'* matrisi *A* matrisinin eşlenik transpozesidir. Modellere yerleştirdiğimiz kavram matrisi, doküman skor matrisi olarak da işlev gören, *U* matrisidir (Yucel, 2016).

2.3. Genelleştirilmiş Lineer Model (Generalized Linear Model)

Y bir bağımlı değişken ve $X_1, X_2, \ldots, X_n$ rastgele değişkenler olsun öyle ki, $P(Y \mid X_1, X_2, \ldots, X_n)$ şartlı dağılımına sahip genelleştirilmiş lineer model, genel olarak şu bileşenlerden oluşur (Levy, 2012):

1- X_i'lerin Y üzerindeki etkileri η fonksiyonu ile ifade edilir öyle ki η X_i'lerden oluşan lineer bir kombinasyondur.
2- Model bir link fonksiyonu üzerine kurulur.
3- $l: \eta \rightarrow E(Y) = \mu$ tersinir bir fonksiyon olmak üzere,

$$\eta = \beta_0 + \sum_{i=1}^{n} \beta_i X_i \quad (10)$$

ifadesi bir genelleştirilmiş lineer modeldir öyle ki $\eta = l(\mu)$ link fonksiyonudur.

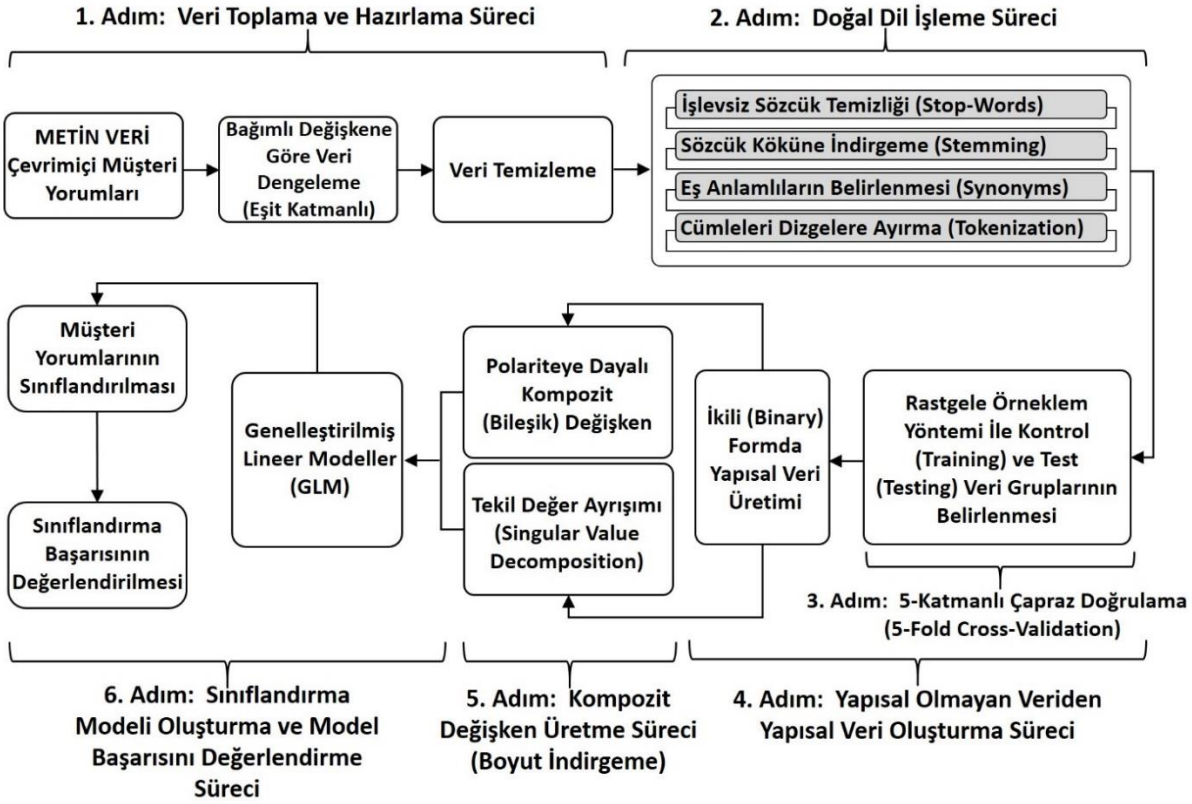
3. Uygulama ve Bulgular (Application and Findings)

3.1. Veri (Data)

Çalışmada kullanılan *verinin* orijinal adı 'Women's Clothing E-Commerce'. Veri ikili (binary) formda olup, internet üzerinden bayan kıyafetleri satışı yapan bir firmanın müşterilerinin, ürünler hakkındaki yorumlarını ve beğenme düzeylerini içermektedir. Verinin bağımlı değişkeni 'Recommend' (tavsiye) (0: No, 1: Yes) olarak belirlenmiştir. Veri içinde toplam 7480 tüketici yorumu bulunmaktadır ve her iki kategori (0/1) için tüketici yorum sayıları eşittir. Veri gerçek bir ticari işletmeleri konu aldığından, veri içinde yer alan firma isimleri 'perakendeci' olarak değiştirilmiştir. Veri CC0 1.0 Universal lisansına sahiptir. Lisansın detaylarına creativecommons.org/publicdomain/zero/1.0/ adresinden ulaşabilirsiniz.

3.1. Uygulama Adımları (Implementation steps)

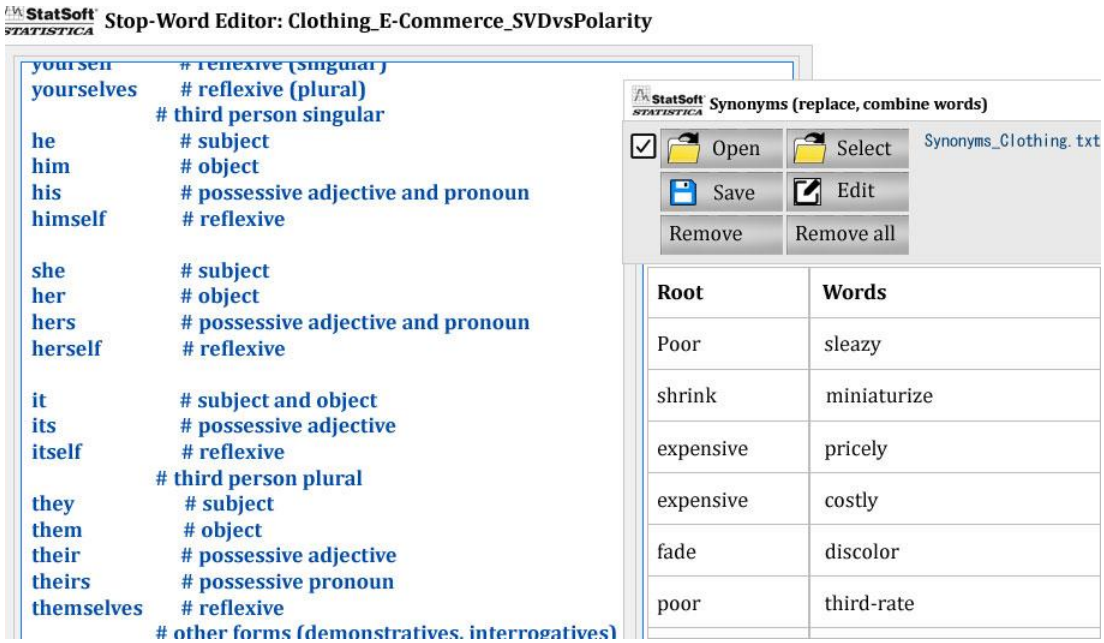
Çalışma 6 temel adımdan oluşmaktadır. Birinci adım üç aşamalı olarak uygulanır. İlk aşama, uyumlu bir algoritma ile direk olarak veri kaynağından veya hazır paket olarak ilgili bir veri tabanından veri teminidir. Çalışmada kullanılan veriseti, bir online veri paylaşım platformu olan Kaggle.com'dan alınmıştır. İkinci aşamada, belirlenen bağımlı değişkenin kategorilerine göre dağılımın dengeli olması için, rastgele seçim yöntemini kullanarak her iki kategorinin de (0 veya 1) eşit sayıda doküman sayısına sahip olması sağlandı. Verinin ham halinde alınan kaynağıyla alakalı olarak, bazı link, dosya uzantıları, adres bilgileri, sayfa/sıra numaraları gibi çalışmayla ilgisi olmayan unsurlar bulunabilir. Bu tarz unsurlar üçüncü aşamada veriden temizlenir ve böylece ilk adım tamamlanmış olur.



Şekil 1: Yöntem ve uygulama adımları(Method and implementation steps)

İkinci adım genel olarak doğal dil işleme süreci şeklinde ifade edilebilir. İlgili veri İngilizce metinlerden oluşmaktadır. Metin işlemede kullanılan metin madenciliği aracı StatSoft Statistica12, İngilizce için hazır ve genel bir 'işlevsiz sözcükler' (stop-words) listesi sağlamaktadır. Mevcut veride tespit edilen diğer

işlevsiz sözcükler de aynı listeye eklenmiştir. Bunun yanında, Statistica İngilizce dili için, kelimeleri köklerine indirgeyen (stemming) ve cümleleri dizgelere ayıran (tokenization) algoritmaları da sağlamaktadır. Ek olarak, veri içinde yer alan eş anlamlı terimler belirlenip, dil analizinden önce programa yüklenmiştir. Şekil 2'de Statistica metin madenciliği sözlük oluşturma sürecine ilişkin bir ekran görüntüsü paylaşılmıştır.



Şekil 2. Metin madenciliği sözlük oluşturma süreci (Text mining dictionary creation process)

Üçüncü adımda, veri eşit sayıda doküman içeren, birbirinden bağımsız ve kesişimi boş küme olan 5 katmana rastgele seçim yöntemiyle ayrılmıştır öyle ki her katmanda yer alan dokümanların, bağımlı değişkenin kategorilerine göre dağılımı da eşit sayıdadır. Şekil 3 'de GLM tanımsal kodlama altında yer alan 'SAMPLE' kodu ile ifade edilen test grupları belirlenmektedir.

Dördüncü adım temel olarak iki aşamadan oluşmaktadır. İlk aşamada, oluşturulan beş katmandan biri (%20) test verisi olarak, geriye kalan dört katman ise kontrol verisi olarak düzenlenmiştir. Bu işlem, her bir katmanın, sırasıyla bir kez test verisi olarak düzenlenmesini ve böylece beş adımda verinin tamamının test edilmesini sağlamak için uygulanmıştır. İkinci aşamada ise, veriden ikili (binary) formda bir matris elde edilmiştir. Bu matris terim-doküman frekans tablosu olarak adlandırılmaktadır. Genel olarak dördüncü adım, yapısal olmayan (metin) veriden, yapısal ve ikili (binary) formda bir veri elde etme süreci olarak ifade edilebilir.

Beşinci adımda, temelinde yukarıda verilen matematiksel formülleri çalıştıran algoritmalar yardımıyla, polariteye dayalı kompozit değişken ve TDA'ya dayalı doküman skorları hesaplanmıştır. Kompozit değişken ve doküman skorlarının bulunduğu değişkenler, sonraki adımda oluşturulacak olan modellerde bağımsız değişken olarak yer almaktadır.

```
Distribution : BINOMIAL
Link function: LOGIT
Response variable : "Recommended{1:Y/O:N)"

Codes of dependent variable
O : Primary code  (1)
1: Secondary code (0)

Design Effects:
Continuous effects : "REVIEW_REC1_RATE_"
Categorical effects:

Model specification:
GLZ ;
RESPONSE = "Recommended{1:Y/O:N)" (0 1) ;
GROUPS = none;
COVAR ATE = ""REVIEW_REC _RATE_1";
DESIGN = ""REVIEW_REC _RATE_1;
INTERCEPT = include :
PARAM = sigma;
SDELTA = 7;
SURFACE = none;
MIXTURE = none;
SAMPLE = "CV_TESTING1" { 102.) ;
COUNTY = none;
MBUILD = all;
CONVERGE = 7:I
MAXITER = 100 ;
IN TIALS   = none;
OFFSET = none;
OUTPUT  = none;
```

Şekil 3. GLM tanımsal kodlama (GLM definitive coding)

Çalışmanın altıncı ve son adımında ise, beşinci adımda üretilen iki grup bağımsız değişkenler yardımıyla Genelleştirilmiş Lineer Modeller kurulmaktadır ve böylece her bir dokümanın bağımlı değişkende karşılık geldiği kategori (0 veya 1) tahmin edilmeye çalışılmıştır. Bir başka ifadeyle, tüketici yorumunun duygu analizi yapılmıştır. Bu amaçla, oluşturulan model kullanılarak tüm dokümanlar 0 ve 1 kategorilerine göre sınıflandırılmıştır. Son aşamada, her iki bağımsız değişken grubuna dayalı yapılan sınıflandırmaların doğruluk oranları hesaplanmıştır ve böylece hem modellerin hem de değişken gruplarının başarıları ölçülmüştür. Şekil 4'te GLM'e ait parametre ve katsayı belirleme detayları verilmiştir.

```
<Header copyright="STATISTICA Data Miner, Copyright (c) StatSoft, Inc., www.statsoft.com."/>
<DataDictionary numberOfFields="2">
        <DataField name="Recommended(1:Y/0:N)" optype="categorical">
                <Value value="0" NumericValue="0"/>
                <Value value="1" NumericValue="1"/>
        </DataField>
        <DataField name="REVIEW_REC1_RATE_1" optype="continuous"/>
</DataDictionary>
<GeneralizedLinearModel
        functionName="classification"
        modelName="Generalized linear regression"
        modelType="generalLinear"
        targetVariableName="Recommended(1:Y/0:N)">
<Extension name="Distribution" value="binomial"/>
<Extension name="LinkFunction" value="logit"/>
<ParameterList>
        <Parameter name="p1" label="Intercept"/>
        <Parameter name="p2" label="REVIEW_REC1_RATE_1"/>
</ParameterList>
<FactorList>
</FactorList>
<CovariateList>
        <Predictor name="REVIEW_REC1_RATE_1"/>
</CovariateList>
<PPMatrix>
        <PPCell value="1" predictorName="REVIEW_REC1_RATE_1" parameterName="p2"/>
</PPMatrix>
<Extension name="CorrectDummyCode" value="1"/>
<Extension name="IncorrectDummyCode" value="-1"/>
<ParamMatrix>
        <PCell targetCategory="0" parameterName="p1" beta="3.16184705908780e+001"/>
        <PCell targetCategory="0" parameterName="p2" beta="-6.45249167608618e+001"/>
</ParamMatrix>
        </GeneralizedLinearModel>
        </PMML>
```

Şekil 4. StatSoft Statistica TMİD (Tahmin Modeli İşaretleme Dili) GLM algoritması (StatSoft Statistica TMİD (predictive model markup language) glm algorithm)

4. Sonuçlar (Conclusions)

Kullanılan veride toplam 7480 tüketici yorumu bulunmaktadır. 5 katmanlı çapraz doğrulama testi için, her bir test grubunda (%20) toplam 1496 tüketici yorumu bulunmaktadır. Tüketici sayıları her iki kategori (0 veya 1) için eşit (748 adet) şekilde dağıtılmıştır. Başka bir ifadeyle, veri ayrık ve eşit dağılıma sahip beş parçaya ayrılmıştır. Bu şekilde uygulanan modelin performansı beş kez test edilmiştir. Lineer modeller binomial dağılıma sahip bir veri üzerine uygulanmıştır ve logit olasılık fonksiyonu kullanılmıştır. Her bir katman için iki kez lineer model oluşturulmuştur. Her iki modelde de bağımlı değişken olarak, tüketici cevap bilgisini içeren '*Recommended* (0/1)' değişkeni kullanılmıştır. Birinci modelde bağımsız değişkenler TDA skorları, ikinci modelde ise kompozit değişkenler olarak belirlenmiştir. Lineer Modellerin tamamı α = 0,05 anlamlılık düzeyinde, istatistiksel olarak anlamlı sonuç vermiştir. Bu süreçte modeller iki aşamalı test edilmiştir. Birinci aşamada p-değer (α = 0,05) parametresinin anlamlı olması beklenmiştir. İkinci aşamada, anlamlı bulunan modelin 'mutlak doğruluk/mutlak doğru tahmin' performansı ölçülmüştür. Bu ölçüm sonuçları karışıklık matrisleriyle ifade edilmiştir. Tablo 1'de beş katman için, modellerin kategorilere göre yaptığı doğru ve yanlış tahmin sayıları (karışıklık matrisleriyle) verilmektedir. Modellerin doğruluk oranları Şekil-5'te verilmiştir.

Tablo 1. Her bir katman için, GLM tahminlerine göre doğru veya yanlış sınıflandırma sayıları (Number of correct or incorrect classifications for each layer based on GLM estimates)

Dokümanların Sınıflandırılması		TDA		POLARİTE	
		Tahmin 0	Tahmin 1	Tahmin 0	Tahmin 1
Katman 1	Gözlem 0	653	95	708	40
	Gözlem 1	131	617	84	664
Katman 2	Gözlem 0	648	100	698	50
	Gözlem 1	133	615	80	668
Katman 3	Gözlem 0	628	120	683	65
	Gözlem 1	127	621	88	660
Katman 4	Gözlem 0	628	120	639	109
	Gözlem 1	144	604	76	672
Katman 5	Gözlem 0	597	151	634	114
	Gözlem 1	149	599	86	662

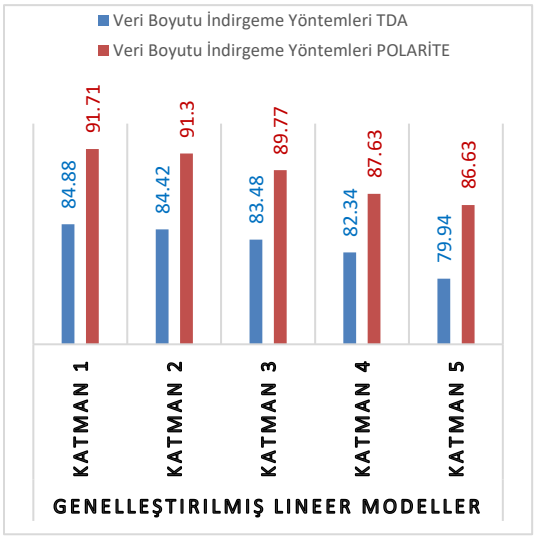
Tablo 2, Tablo 1'de verilen doğru tahmin sayılarının toplam tahmin sayılarına oranı, başka bir ifadeyle, modellerin başarı oranları verilmektedir. Sonuçlardan da açıkça görüldüğü gibi, polariteye dayalı kompozit değişken tüm katmanlarda ortalama %6 daha iyi sonuç vermektedir. Polariteye dayalı değişkenin modellerde sağladığı başarı oranı en düşük %86,63 ve en yüksek %91,71 arasında değişmektedir. Bu durum son yıllarda yapılan benzer polarite çalışmaları dikkate alındığında oldukça dikkat çekici bir başarı sunmaktadır.

Tablo 2. Modellerin doğruluk oranları (The accuracy of the models)

Doğruluk Oranları (%)		Veri Boyutu İndirgeme Yöntemleri	
		TDA	POLARİTE
Genelleştirilmiş Lineer Modeller	Katman 1	84,88	91,71
	Katman 2	84,42	91,30
	Katman 3	83,48	89,77
	Katman 4	82,34	87,63
	Katman 5	79,94	86,63

İlgili çalışmaların anlatıldığı kısımda da ifade edildiği üzere, *Zubrinic'*in 2018'de yayınladığı *tüketici*

Duygu analizi çalışmasının, ikili (binary) bir veride elde ettiği doğruluk oranı maksimum %84,5 civarında kalmıştır (Zubrinic, 2018). Bunun yanında *Zhao* ve *Xu'*nun detaylarının yukarıda verildiği, 2018'de yayınlanan *fikir tespiti* (opinion detection) çalışmasında, modellerin doğruluk oranları en iyi %70,02 civarında kalmıştır (Zhao, 2018).



Şekil 5. Modellerin doğruluk oranlar (The accuracy of the models)

Kaynaklar (References)

Al-Otaibi, S., Alnassar, A., Alshahrani, A., Al-Mubarak, A., Albugami, S., Almutiri, N., Albugami, A., 2018. Customer Satisfaction Measurement Using Sentiment Analysis, International Journal of Advanced ComputerScienceand Applications (IJACSA), Vol.9, No.2.

Arunachalam, N., Sneka S. J., Mathi, G. M., 2017. A Survey On Text Classification Techniques For Sentiment Polarity Detection, Innovations in Powerand Advanced Computing Technologies (i-PACT), 1-5. 10.1109/IPACT.2017.8245127.

Boling, C., Das K., 2015. Reducing Dimensionality of Text Documents Using Latent Semantic Analysis, International Journal of Computer Applications (0975 – 8887), Vol.112, No.5.

Levy, R., 2012. Probabilistic Models in the Study of Language , ch. 6, pp: 107-108.

Pajupuu, H., Altrov, R., Pajupuu, J., 2016. Identifying Polarity in Different Text Types, pp 126-138, doi.org/10.7592/FEJF2016.64.polarity.

Pipino, L. L., Lee, Y. W., Wang, R. Y., 2002. Data Quality Assessment, Communications Of The ACM, Vol.45.

Rajalakshmi, Narayanan, M., Ramkumar, M., 2015. An Exclusive Study on Unstructured Data Mining with Big Data, International Journal of Applied Engineering Research, Vol.10, No.4, pp.3875-3886.

Singh, V. K., Piryani, R., Waila, P., Devaraj, M., 2014. Computing Sentiment Polarity of Texts at Document and Aspect Levels, ECTI transactions on computer and information technology, Vol.8, No.1.

Tian, Y.,Stewart, C., 2008. History of E-Commerce, DOI:10.4018/978-1-59904-943-4, ch001.

Tomar, D. S., Sharma, P., 2016. A Text Polarity Analysis Using Sentiwordnet Based an Algorithm, International Journal of ComputerScienceand Information Technologies (IJCSIT), pp. 190-193, Vol.7.

Varghese, N., Verghese, V., Gayathri, P., Jaisankar, N., 2012. A Survey Of Dimensionality Reduction And Classification Methods, International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.3.

Yom-Tov, G. B., Ashtar, S., Altman, D., Natapov, M., Barkay, N., Westphal, M., Rafaeli, A., 2018. Customer Sentiment in Web-Based Service Interactions: Automated Analyses and New Insights, International World Wide Web Conference Committee (IW3C2), Creative Commons CC BY 4.0 License.

Yucel, A., 2016. Predictive Text Analytics And Text Classification Algorithms, A dissertation submitted to the Graduate Faculty of Auburn University, pp 19.

Zhao, S., Xu Z., Liu L., Guo M., Yun, J., 2018. Towards Accurate Deceptive Opinions Detection Based on Word Order-Preserving CNN, Mathematical Problems in Engineering, Article ID 2410206, Vol. 2018

Zhenxiang, W., Lijie, Z., 2011. Case Study of Online Retailing Fast Fashion Industry, International Journal of e-Education, e-Business, e-Management and e-Learning, Vol.1, No.3.

Zubrinic, K., Milicevic, M., Sjekavica, T., 2018. Obradovic, I., A Comparison of Machine Learning Algorithms in Opinion Polarity Classification of Customer Reviews, International Journal of Computers, Vol.3.