Sentiment Analizinde Öznitelik Düşürme Yöntemlerinin Oto Kodlayıcılı Derin Öğrenme Makinaları ile Karşılaştırılması

Oğuz KAYNAR[1], Zafer AYDIN[2], Yasin GÖRMEZ[3]

[1]Yönetim Bilişim Sistemleri Bölümü, Cumhuriyet Üniversitesi, Sivas, Türkiye
[2]Bilgisayar Mühendisliği Bölümü, Abdullah Gül Üniversitesi, Kayseri, Türkiye
[3] Yönetim Bilişim Sistemleri Bölümü, Cumhuriyet Üniversitesi, Sivas, Türkiye
okaynar@cumhuriyet.edu.tr, zafer.aydin@agu.edu.tr, yasingormez@cumhuriyet.edu.tr
(Geliş/Received:06.03.2017; Kabul/Accepted:15.07.2017)
DOI: 10.17671/gazibtd.331046

Özet— Günümüz teknolojisinde internetin her kesim tarafından çok yoğun olarak kullanılmasından dolayı insanlar artık görüş, fikir ve hislerini sosyal paylaşım siteleri, forum, blog benzeri birçok ortam aracılığı ile paylaşmaya başlamıştır. Ancak her geçen gün artan veri sayısı ve boyutu, bu verilerden manuel olarak anlamlı bilgiler çıkartılmasını çok zahmetli ve pahalı bir iş haline getirmektedir. Otomatik olarak verinin duygu içerip içermediğinin saptanması ve bu duygunun olumlu, olumsuz veya tarafsız olma durumunun belirlenmesi duygu analizi yardımıyla gerçekleştirilmektedir. Duygu düşünce analizinde, konuşma dilinin karmaşıklığı, değerlendirilen metin sayısının fazlalığı ve uzunluğu, çok sayıda gereksiz ve gürültü içeren öznitelik vektörüne neden olmaktadır. Boyut problemi olarak adlandırılan bu durum hesaplama zamanın artmasına ve sınıflama hatalarına yol açmaktadır. Bu çalışmada ise bahsedilen problemlere çözüm olarak önerilen derin öğrenme tabanlı oto kodlayıcı (Autoencoder) modeli ile gürültü giderici oto kodlayıcı (Denoising Autoencoder) modeli boyut düşürme tekniği olarak kullanılmış ve literatürde yaygın olarak kullanılan diğer boyut düşürme teknikleri ile kıyaslanmıştır. Elde edilen tüm veri setleri için sınıflama algoritması olarak Destek Vektör Makinaları ve Yapay Sinir Ağları kullanan farklı modeller geliştirilmiştir. Yapılan analizlerin sonucunda, boyut düşürme tekniklerinin duygu analizi için elde edilen sonuçları iyileştirdiği, önerilen oto kodlayıcı modellerinin ise var olan tekniklere benzer ya da onlardan daha iyi sonuçlar aldığı gözlemlenmiştir.

Anahtar Kelimeler— Boyut düşürme, Oto kodlayıcı, Yapay sinir ağları, Destek vektör makineleri, Duygu analizi, Derin öğrenme

Comparison of Feature Reduction Methods with Deep Autoencoder Machine Learning in Sentiment Analysis

Abstract— Because the internet is extensively used by people from all strata with today's technology, people now share their opinions, ideas and feelings through a variety of media such as social networking sites, forums and blogs. However, the number and size of data that is increasing day by day makes it very laborious and expensive to extract meaningful information manually from these data. Determination of whether data includes emotions or not automatically and determination of these feelings being positive, negative and neutral are performed by sentiment analysis. In sentiment analysis, the complexity of the speech language, the excessive number and length of texts being evaluated causes a large number of unnecessary and noise-containing feature vectors. This situation, which is called dimensionality problem, leads to increase of computation time and classification errors. In this study, a deep autoencoder model and a denoising autoencoder model are proposed and used as dimension reduction methods to overcome mentioned problems and compared with other feature reduction methods commonly used in literature. For all data sets obtained, different models have been developed using Support Vector Machines and Artificial Neural Networks as the classification algorithm. According to the analyses made, it has been observed that the feature reduction methods improve the results obtained of sentiment analysis, and the proposed autoencoder models have similar or better results than the existing methods.

Keywords— Feature Reduction, Autoencoder, Artificial Neural Networks, Support vector machines, Sentiment Analysis, Deep learning

1. GİRİŞ (INTRODUCTION)

Günümüzde teknoloji alanındaki gelişmeler ile birlikte internet; sağlık, bilim, eğlence, spor, sanat gibi insan hayatının hemen hemen her alanına girmeyi başarmıştır. Geliştirilen web sayfaları ve web uygulamaları sayesinde bu alanlardaki yorum, fikir ya da düşünceler rahatlıkla paylaşılabilmektedir. Bu paylaşımlar ile birlikte internet devasa büyüklükte bir metin deposu haline dönüşmüştür. İnternet ortamı bu özelliğinden dolayı aranan birçok bilgiyi barındırmasına rağmen metin sayısındaki bu fazlalık, doğru bilgiye ulaşmayı karmaşık ve güç hale getirmektedir. Bu nedenle son yılların popüler konularından metin madenciliği uygulamaları, internet siteleri üzerinde de sıkça kullanılmaya başlanmıştır. Metin madenciliği özetle istatistiksel yöntemler ve makine öğrenmesi yöntemleri yardımı ile metinden anlamlı ve kullanılabilir bilgilerin elde edilme süreci olarak tanımlanabilir. Metin madenciliği; metinlerin özetlenmesi, sınıflandırılması, kümelenmesi, bilgi çıkarımı ve metinlerde geçen duygu ve düşüncelerin analizi gibi daha birçok alt dala bölünebilmektedir. Bu alt dallardan duygu ve düşünce analizi ise, metindeki fikrin, görüşün ya da duygu durumunun matematiksel modeller yardımı ile çıkarılması işlemidir.

Duygu analizi, sözlüğe dayalı modeller ve makine öğrenmesine dayalı modeller olmak üzere ikiye ayrılmaktadır. Sözlüğe dayalı modellerde, ilk olarak metinlerde hangi duygu durumlarının aranmak istendiği belirlenir. Daha sonra belirlenen duygu durumlarını ifade eden kelimeler ve o kelimelerin anlamdaşları metin içerisinde aranarak her bir kelime için bir sözlük yardımıyla duygu durumu gösteren bir skor değeri elde edilir. Son adımda ise istatistiksel yöntemler ile metnin hangi duygu durumunu ifade ettiği tahmin edilir. Makine öğrenmesine dayalı yöntemlerde ise, ilk olarak metinler etiketlenir. Ardından bu metinler çeşitli metin madenciliği yöntemleri ile temizlenerek ön işlemeden geçirildikten sonra sınıflandırmaya uygun hale getirilmek üzere vektör uzay modelleri oluşturulur. Ardından bu vektör uzayları tercihe göre eğitim, test, doğrulama gibi alt setlere bölünür. Son olarak model, eğitim ve doğrulama veri setleri yardımı ile eğitilir ve test verileri yardımıyla duygu durumu tahmini yapılır. Sözlüğe dayalı yöntemlerde her kelime için pozitif, negatif ve nötr ağırlık skorlarını içeren önceden tanımlı sözlüğe ihtiyaç duyulmakta ve bu sözlük her dil için henüz bulunmamaktadır. Makine öğrenmesi yöntemlerinin dilden bağımsız olması ve yüksek başarı oranları elde etmesi akademik alanda duygu analizi için daha çok tercih edilmesine neden olmuş ve bu amaçla birçok sistem tasarlanmıştır.

Makine öğrenmesi yöntemleri, denetimli ve denetimsiz öğrenme olarak iki ana başlık altında toplanmaktadır. Denetimli öğrenmeyi denetimsiz öğrenmeden ayıran en önemli özellik eğitim esnasında veri setinin etiket bilgisinden yararlanıyor olmasıdır. Duygu analizi ile ilgili yapılan çalışmalar incelediğinde denetimli makine öğrenmesi yöntemlerinin daha çok tercih edildiği görülmektedir. Liu ve diğerleri ortaya atmış oldukları Çin karakter tabanlı bigram öznitelik çıkarma yöntemini, literatürde yer alan bigram, trigram ve kelime tabanlı unigram yöntemleri ile karşılaştırmak için denetimli makine öğrenmesi yöntemlerinden Destek Vektör Makinaları (DVM), Naive Bayes (NB) ve Yapay Sinir Ağları (YSA) ile Çin web sitelerinden elde edilen 16000 metni kullanarak farklı modeller tasarlamışlar. Önermiş oldukları öznitelik çıkarma yöntemi diğer yöntemlere üstünlük sağlayarak DVM ile birlikte kullanılan modelde %91,62 oranında F1 skoru elde edilmişti [1]. Alec vd. NB, DVM ve Maksimum Entropi (ME) yöntemlerini kullanan üç farklı modeli twitter verisi kullanarak eğitmiş ve %83 başarı oranı elde etmişlerdir [2]. Mouthami vd. önerdikleri bulanık mantık yöntemini Cornell film verisi üzerinde test ederek başarı oranını yükseltmişlerdir [3]. Singh vd. NB, DVM, modifiye edilmiş Wordnet ve semantik yönlendirme yaklaşımı yöntemleri ile üç farklı veri seti kullanarak tasarladıkları modellerde %88,8'e varan başarı oranı elde etmişlerdir [4]. Gautham ve Yadav NB, DVM, ME ve Wordnet yaklaşımları ile tasarlamış oldukları modellerde %83,8 ile %89,9 arasında başarı oranı elde etmişlerdir [5]. Nizam ve Akın dengeli ve dengesiz veri seti kullanmanın başarı oranına etkisini göstermek için twitter verileri ile iki farklı set oluşturmuş; oluşturulan bu veri setlerini NB, Rastgele Orman (RO), Sıralı Minimum Optimizasyonu, J48 ve K en yakın komşu (Knn) algoritmalarını kullanarak sınıflamış ve dengeli veri seti için %6'lara varan daha iyi başarı oranı elde etmişlerdir [6]. Çoban vd. Türkçe twitter verilerini kullanarak eğittikleri NB, Multinomial Naive Bayes (MNB), DVM ve KNN modellerinde %66,06 başarı oranı elde etmişlerdir [7]. Kranjc vd. aktif öğrenmeye dayalı yöntemin duygu analizindeki etkisini test etmek için DVM ile iki farklı model oluşturmuş ve aktif öğrenmeye dayalı modelin %6,7 daha başarılı olduğunu gözlemlemişlerdir [8]. Tripathy vd. n-gram özellik çıkarma yöntemini dört farklı sınıflama algoritması kullanarak denemiş ve %95 başarı oranı elde etmişlerdir [9]. Rohini vd. duygu analizinde metin dilinin önemini göstermek için İngilizce yazılmış metinler ile Kannada dilinde yazılmış metinleri kıyaslayarak, İngilizce yazılan metinlerin daha başarılı sonuçlar verdiğini göstermişlerdir [10].

Sınıflama yöntemlerini tek başına kullanarak bir model geliştirmek mümkün olduğu gibi, yöntemin kendine has dezavantajlarından kaynaklanan hatalarını gidermek amacı ile başka yöntemler ile birlikte kullanılması da mümkündür. Ensemble adı verilen bu yöntemlerde iki ya da daha fazla sınıflama algoritmasının sonuçları, içlerinden biri temel algoritma olmak üzere, birleştirilerek nihai sonuç elde edilir. Xia vd. DVM, NB, ME algoritmalarını üç farklı ensemble yöntemi ile birleştirmiş; Cornell film verisinden elde ettikleri üç farklı veri seti ile modelleri eğiterek %81,12 oranında başarı sağlamışlardır [11]. Neethu ve Rajasree DVM, ME ve NB algoritmalarını twitter verisi üzerinde ensemble yöntemi ile birleştirmiş ve %90 başarı oranı elde etmişlerdir [12]. Fersini vd. Bayes tabanlı ensemble yöntemini diğer

ensemble yöntemleri ile kıyaslamak için NB, ME, DVM ve Koşullu Rasgele Alanlar sınıflama algoritmalarını kullanarak altı farklı veri seti ile analiz yapmışlardır. Bu analizin sonucunda Bayes tabanlı yöntemin başarı oranını artırdığı ve zaman kazandırdığı görülmüştür [13]. Da Silva vd. MNB, DVM, RO ve Lojistik Regresyon (LR) algoritmalarını, önermiş oldukları ensemble yöntemi ile birleştirerek beş farklı veri seti üzerinde %76,84 ile %87,20 arasında başarı oranı elde etmişlerdir [14]. Çatal ve Nangir NB ve DVM algoritmalarını çeşitli ensemble yöntemleri ile birleştirerek %86,13'e varan başarı oranı elde etmişlerdir [15].

Makine öğrenmesi yöntemlerinde sınıflama algoritmaları kadar, veri setinden elde edilen özniteliklerin kalitesi de başarı oranına etki eden önemli bir unsurdur. Eğitim sırasında sınıflandırma performansını olumsuz yönde etkileyen gereksiz ve gürültü içeren öznitelikleri elemek amacıyla veri setleri üzerinde çeşitli boyut düşürme ve öznitelik seçim yöntemleri sıkça kullanılmaya başlanmıştır. Tan ve Zhang veri seti üzerinde Doküman Frekansı (DF), Ki Kare (Chi), Bilgi Kazancı (BK) ve Karşılıklı Bilgilendirme (KB) öznitelik seçim yöntemlerini uygulamış; elde edilen yeni veri setleri ile beş farklı sınıflama algoritmasını kullanarak eğitmiş oldukları modellerde %88,58'e varan başarı oranı elde etmişlerdir [16]. Go vd. Chi, KB, ME (Maksimum Entropy) ve Frekans Tabanlı öznitelik seçim yöntemlerini twitter verisi üzerinde uygulayarak %84 başarı oranı elde etmişlerdir [17]. Meral ve Diri korelasyon tabanlı özellik seçimi yöntemlerini NB, DVM, RO sınıflama algoritmaları ile birlikte kullanarak twitter verisi üzerinde %90 F1 skoru elde etmişlerdir [18]. Vinodhini ve Chandrasekaran Temel Bileşen Analizi (TBA) boyut düşürme yöntemini NB ve DVM sınıflama algoritmaları ile birlikte kullanarak %77 başarı oranı elde etmişlerdir [19]. Yousefpour vd. SVM, NB, ME algoritmaları ve bunların ensemble yönteminden oluşan dört farklı modeli, iki farklı boyut düşürme algoritması ile birlikte kullanarak yapmış oldukları uygulamalarda %90,91 başarı oranı elde etmişlerdir [20]. Kim ve Lee önerdikleri yarı denetimli boyut düşürme yöntemini dört farklı veri setine uygulamışlar. Boyut düşürme sonucu elde edilen yeni veri setlerini iki farklı sınıflama algoritmasıyla test ederek önerdikleri boyut düşürme yöntemini diğer boyut düşürme yöntemleriyle kıyaslamışlar, önerilen bu yöntemin daha başarılı sonuçlar verdiğini belirtmişlerdir [21]. Shyamasundar ve Rani twitter verisi üzerinde boyut düşürme yöntemlerini uygulamış ve orijinal veri seti kullanılarak elde edilen sonuçlar ile boyutu düşürülmüş veri seti kullanılarak elde edilen sonuçları kıyaslamışlardır. Her ne kadar boyutu düşürülmüş veri seti ile daha kötü sonuçlar elde etmiş olsalar da, bu sonuçların boyut düşürülmemiş veri seti ile elde edilen sonuçlara yakın olması ve veri setindeki küçülme nedeni ile bu yöntemlerin kullanılmasını önermişlerdir [22].

Makine öğrenmesine dayalı duygu analizi yöntemlerinde ilk adım olan öznitelik çıkarma işlemi için; Bigram, Unigram, Trigram ve bunların çeşitli kombinasyonları gibi birçok yöntem bulunmaktadır. Metin sayısı ve uzunlukları arttıkça çıkarılan özniteliklerin sayısı da artmaktadır. Metinler birbirinden farklı birçok kelime içerdiği için çıkarılan vektör uzayları çok sayıda sıfır içermektedir ve her bir metin için bu sıfırlar vektör uzayında farklı indislerde yer almaktadır. Vektör uzaylarındaki bu durumun boyut çıkmazı (curse of dimensionality) ve seyreklik (sparsity) problemlerini beraberinde getirmesi ise duygu analizinde başarı oranının düşmesine neden olmaktadır. Sözü edilen sorunların üstesinden gelmek için boyut düşürme yöntemleri sıklıkla kullanılmaktadır. Boyut düşürme, elde edilen vektör uzaylarında çeşitli matematiksel, istatiksel ya da makine öğrenmesine dayalı yöntemler uygulanarak orijinal veri setinin daha küçük bir boyutta mümkün olduğu kadar az bilgi kaybı ile ifade edilmesi işlemidir. Veri setinin boyutunun azalmasıyla birlikte eğitim sırasında gereksinim duyulan bellek miktarı da azalarak, öğrenme işleminin daha etkin ve daha doğru gerçekleşmesi sağlanmaktadır.

Bu çalışmada boyut düşürme için önerilen iki farklı tip oto kodlayıcı tabanlı derin öğrenme makinesi ve literatürde yaygın olarak kullanılan Temel Bileşen Analizi (TBA), Çekirdek Tabanlı Temel Bileşen Analizi (ÇTBA), Tekil Değer Ayrışımı (TDA), Faktör Analizi (FA) gibi boyut düşürme yöntemleri kullanılmıştır. Elde edilen yeni veri setleri ile boyut düşürülmemiş veri seti üzerinde DVM ve YSA gibi iki farklı sınıflama algoritması kullanılarak sonuçlar karşılaştırılmıştır. Sınıflama algoritmasından doğacak farklılıkların boyut düşürme tekniklerine etki etmemesi için bu teknikler iki sınıflama algoritması için de ayrı ayrı değerlendirilmiştir. İkinci bölümde, çalışmada kullanılan boyut düşürme ve sınıflama yöntemlerinin teorik çerçevesi verilmiş, üçüncü bölümde uygulama gerçekleştirilmiş ve elde edilen sonuçlar karşılaştırılmış, son bölümde ise değerlendirme ve önerilerde bulunulmuştur.

2. YÖNTEMLER (TECHNIQUES)

2.1. Öznitelik Düşürme Teknikleri (Feature Reduction Methods)

2.1.1. Tekil Değer Ayrışımı (Singular Value Decomposition)

Boyut düşürme yöntemi olarak da kullanılabilen tekil değer ayrışımı (TDA) bir matrisi çarpanlarına ayıran önemli bir lineer cebir yöntemidir. Bu çarpanlarına ayırma işlemindeki temel amaç matristeki satır ve sütunların bağımlılıkları dikkate alınarak matrisi temsil etme değerlerinin elde edilmesidir. Bu amaç doğrultusunda $A_{m*n} = U_{m*r} \times S_{r*r} \times (V_{n*r})^t$ eşitliği sağlanacak şekilde, m satır ve n sütuna sahip A matrisi; m satır ve r sütunu olan U, r satır ve r sütunu olan S ve n satır ve r sütunu olan V matrisleri olmak üzere üç matrise ayrılır. U matrisi sol tekil vektörler, S matrisi tekil değerler ve V matrisi de sağ tekil vektörler olarak adlandırılmaktadır. S matrisi, azalan sıra ile sütunların temsil değerlerinin bulunduğu bir köşegen matrisidir. Bu eşitlikte daha önce de belirtildiği gibi m ve n orijinal

matrisin satır ve sütunlarını temsil ederken, r sayısı S matrisinin rank değerini temsil etmektedir ve bu sayı n değerine ya eşit ya da ondan daha küçüktür. Bu rank değerlerine bakılarak, belirlenen eşik değerinin altında kalan sütunlar göz ardı edilerek bulunmuş olan yeni üç matris yardımı ile m*d (d<r) boyutlarında yeni A matrisi elde etmek mümkündür. TDA bu özelliği ile çok karmaşık olmamasına rağmen güçlü bir boyut düşürme tekniği olarak kullanılabilmekte ve birçok problemde başarı oranlarını yükseltmektedir.

2.1.2. Temel Bileşen Analizi (Principal Component Analysis)

Temel bileşen analizi (TBA), TDA yönteminde olduğu gibi değişkenler arasındaki bağımlılığın bulunması için kullanılan diğer bir boyut düşürme tekniğidir. En büyük varyans ile en az kaybı hedefleyen bu yöntemin ilk aşamasında her bir değişken için diğer değişkenlerle olan kovaryans değeri eşitlik 1'de formülize edildiği gibi hesaplanır. Kovaryans değeri iki değişkenin birlikte değişimini temsil eden bir değerdir. Bu değerin pozitif olması iki değişkenin aynı anda büyüdüğü ya da küçüldüğü, negatif olması iki değişkenden biri büyür iken diğerinin küçüldüğü, sıfır olması ise bu iki değişkenin birbirinden bağımsız olduğu durumu belirtir.

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})\times(Y_i-\bar{Y})}{n-1} \quad (1)$$

Bu eşitlikte X,Y her bir değişkeni; n örnek sayısını; X_i, Y_i ilgili değişken için i'ninci örnek değerini; $\bar{X}$, $\bar{Y}$ ilgili değişken için verilen örneklerin ortalama değerini temsil etmektedir. Bu adımın devamında bulunan kovaryans değerleri kullanılarak kovaryans matrisi oluşturulur. Sonraki adımda ise bu matris kullanılarak özdeğerler ve özvektörler hesaplanır. Hesaplanan bu özvektörler yüksek değerden küçük değere doğru sıralanarak özellik matrisi elde edilir. Bu sıralamadaki amaç, değişkenleri, veri setini temsil kapasitesine göre sıralamaktır. Son adımda ise bu sıralama doğrultusunda istenilen sayıda alınan temsil değeri en yüksek olan bileşenler seçilerek elde edilen matris ile veri seti çarpılır ve boyutu düşürülmüş veri seti elde edilir.

2.1.3. Çekirdek Tabanlı Temel Bileşen Analizi (Kernel Principal Component Analysis)

TBA doğrusal olarak ayrışabilen veri setleri için güçlü bir boyut düşürme tekniğidir ancak doğrusal olarak ayrışamayan veri setleri için başarılı sonuçlar alınamayan durumlar vardır. Bu tip durumlarda veri setlerini daha yüksek bir boyuta taşımak, o veri setini doğrusal olarak ayrılabilir bir duruma getirebilmektedir. Çekirdek Tabanlı Temel Bileşen Analizi (çTBA), doğrusal olarak ayrışamayan veriler için güçlü bir boyut düşürme tekniğidir. Bu yöntemin ilk aşamasında çeşitli çekirdek fonksiyonları kullanılarak değişkenlerin veri seti eşitlik 2'de formüle edildiği gibi bir üst boyuta yeniden haritalanır.

$$K(X_i, X_j) = \Phi(X_i)\Phi(X_j)^T \quad (2)$$

Eşitlik 2'de X_i, X_j; i'ninci ve j'ninci değişkeni, Φ ise kullanılan çekirdek fonksiyonunu (en çok kullanılan üç çekirdek fonksiyonu: Linear, Gaussian, Polynomial) temsil etmektedir. Bir sonraki adımda haritalanan veri seti kullanılarak özdeğerler, özvektörler ve kovaryans matrisi hesaplanıp TBA'ya benzer bir şekilde boyutu düşürülmüş veri seti elde edilir ve model sonlandırılır.

2.1.4. Faktör Analizi (Factor Analysis)

Faktör analizi (FA), değişkenler arasında gözlemlenemeyen gizli bağımlılıkları ortaya çıkarmayı hedefleyen bir boyut düşürme tekniğidir. TBA'dan farklı olarak değişkenlerin birlikte değişiminin gücü ile de ilgilenir. Bu nedenle FA tekniğinde kovaryans matrisi değil, eşitlik 3'te formüle edilen korelasyon değerleri ile elde edilen korelasyon matrisi kullanılmaktadır.

$$Korelasyon\,(X,Y) = \frac{cov(X,Y)}{S_X \times S_Y} \quad (3)$$

Bu eşitlikte $cov(X,Y)$; X ve Y arasındaki kovaryans değerini, S_X; X değişkeninin standart sapma değerini , S_Y ise Y değişkenin standart sapma değerini temsil etmektedir. Korelasyon matrisi, veri setinin FA için uygun olup olmadığını gösteren önemli bir etkendir. Bu nedenle en az iki değişken arasında yüksek korelasyon katsayısı elde edilmelidir. Bu adımdan sonra faktör sayısına karar verilerek ve eşitlik 4 sağlanacak şekilde faktörler tespit edilip boyutu düşürülmüş yeni veri seti elde edilir.

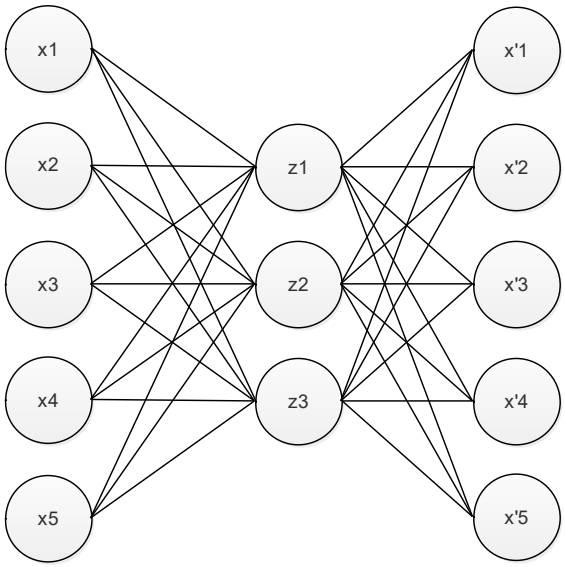
$$X_{n*d} = Y_{n*f} \times W_{f*d} + E_{n*d} \quad (4)$$

Bu eşitlikte X, n örnek ve d özniteliği olan orijinal veri setini; Y, aynı örnek sayısına sahip f özniteliği olan boyutu düşürülmüş veri setini; W, faktörler matrisini; E ise her değişken için sabit katsayılar matrisini temsil etmektedir.

2.1.5. Derin Öğrenme Tabanlı Oto Kodlayıcı (Deep Autoencoders)

Yapay sinir ağlarının bir türevi olan oto kodlayıcı, ilk olarak 1990'lı yıllarda Hinton ve PDB grup tarafından ortaya atılmış, 2006 yılında derin öğrenme mimarisinin güncellik kazanması ile makine öğrenmesindeki ana konulardan biri haline gelmiştir [23]. Oto kodlayıcı, girdi katmanı, gizli katman ve çıktı katmanı olmak üzere üç katmandan meydana gelen tam bağlı bir yapay sinir ağıdır. Ağın giriş çıkışında kullanılan veri seti aynı olduğundan girdi katmanı ve çıktı katmanındaki nöron sayıları birbirine eşit ve veri setindeki öznitelik sayısı kadardır. Gizli katmandaki nöron sayısı ise istenilen şekilde belirlenebilmektedir ve bu sayı ağın performansını etkileyen önemli bir unsurdur. 5 özniteliği olan bir veri

seti için gizli katmanda 3 nöron olan örnek bir oto kodlayıcı mimarisi şekil 1'de gösterilmiştir.



Şekil 1. Oto Kodlayıcı Mimarisi
(Autoencoder Architecture)

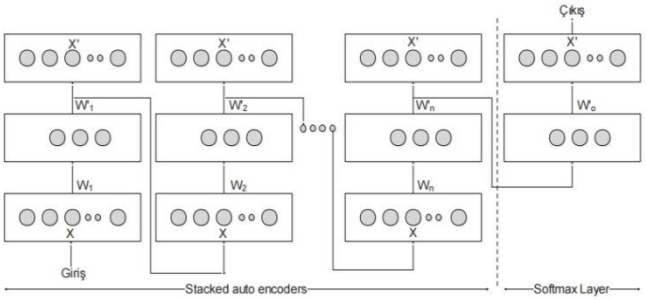
Oto kodlayıcı, ağın çıkışında girdi veri setinin aynısını kullandığından dolayı veri setindeki sınıf bilgisi içeren etiketlere ihtiyaç duymaz, yani denetimsiz bir makine öğrenmesi yöntemidir. Ağ, giriş veri setini çıkışta verilen aynı veri setine uydurmak amacıyla eğitim sırasında geri yayılım algoritmasını kullanarak en uygun ağırlık değerlerini belirler. Bu nedenle yöntem öğreticisi olmayan geri yayılım algoritması olarak da anılmaktadır [24]. Oto kodlayıcı, kendi giriş verisini en az kayıpla yine kendine haritalayan bir kodlayıcı gibi çalışmaktadır. Orta katmanda, giriş ve çıkış katmanından daha az sayıda nöron kullanılması durumunda, boyutu düşürülen veri orta katmanın çıkışından elde edilir. Temel amacı en az kayıpla boyutu en aza indirgemek olan bu modelde, ilk olarak okunan veri eşitlik 5'te gösterilen formül ile girdi katmanından gizli katmana, gizli katmandan da çıktı katmanına aktarılır. Daha sonra eşitlik 6'da gösterildiği gibi gerçek değerler ile ağın hesapladığı çıktı değerleri arasındaki farkın karesini minimize edecek şekilde ağırlıklar güncellenir.

$$y_j = f\left(\sum_{i=1}^{n} x_i \times w_{ij} \right) + b \quad (5)$$

$$\min \sum_{i=1}^{n} \left(x_j' - y_j \right) \quad (6)$$

Eşitlik 5'te x_i, o anki katmanın i'ninci nöronun değerini; y_j, bir sonraki katmandaki j'ninci nöronun değerini; w_{ij}, x_i ile y_j'yi bağlayan ağırlık değerini; n, mevcut katmanın nöron sayısını; b her katman için sabit bias değerini; f ise kullanılan aktivasyon fonksiyonunu (gauss, softmax, sigmoid gibi) temsil etmektedir. Eşitlik 6'da x_j' değerleri gerçek değerleri gösterirken; y_j değerleri ağ tarafından hesaplanan değerleri temsil etmektedir.

Derin oto kodlayıcı ise birçok oto kodlayıcının birbiri ardına katmanlı şekilde bağlanmasıyla elde edilir. Şekil 2'de gösterildiği gibi bir oto kodlayıcı modelinin gizli katmanından elde edilen değerler bir sonraki oto kodlayıcı modelinin girdi katmanı olacak şekilde arka arkaya bağlanır ve benzer şekilde eğitilir. Bu yöntemde her bir oto kodlayıcı modelinin birbirinden bağımsız olarak eğitilmesi bazen başarı oranının düşmesine neden olmaktadır. Bu problem hassas ayar (fine tuning) denilen yöntem ile çözülebilmektedir. Bu yöntemde elde edilen her bir model birleştirilir ve veri etiketleri de dikkate alınarak ağırlıklar topluca yeniden güncellenir.



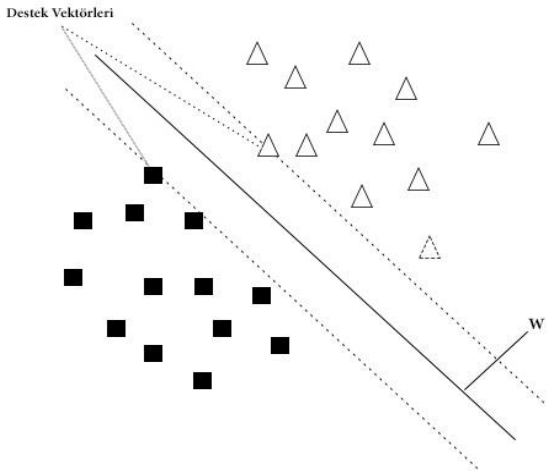
Şekil 2. Derin Oto Kodlayıcı Mimarisi
(Deep Autoencoder Architecture)

Oto kodlayıcı modelindeki çıktı katmanında, girdi katmanındaki verinin aynısını elde etmeye çalışmak, bazı sistemler için o verinin ezberlenmesine ve test verisi için kötü sonuçlar elde edilmesine neden olmaktadır. Bu probleme çözüm olarak üretilen denoising (gürültü giderici) oto kodlayıcı modelinde, girdi katmanında kullanılan veriye çeşitli gürültüler eklenerek, çıktı katmanında gürültüsüz girdi değerleri elde edilmeye çalışılmakta böylece sistem eğitim verisinde olmayan farklı örüntüye sahip verileri de öğrenebilmektedir.

2.2. Sınıflama Algoritmaları (Classification Algorithms)

2.2.1. Destek Vektör Makineleri (Support Vector Machines)

Destek Vektör Makinaları (DVM); doğru, düzlem ya da hiper düzlem yardımı ile verilerin iki sınıfa ayrıldığı bir makine öğrenmesi yöntemidir. Doğrusal olarak ayrışabilen veriler için sıkça kullanılan bu yöntem, çekirdek fonksiyonları yardımı ile verileri doğrusal olarak ayrıştırılabilir duruma getirebildiği için doğrusal olarak ayrışamayan veriler için de kullanılabilmektedir. Bu yöntemdeki asıl amaç, hataların karesini en aza indirecek ayracı belirlemektir. Şekil 3'te görüldüğü gibi hatayı en az yapan ve birbirine paralel iki destek vektörü seçilerek, bu düzlemler arasındaki uzaklık maksimum yapılır. Daha sonra bu iki vektörün orta noktasındaki w vektörü seçilir. Son adımda ise yeni x için $y = w^t x + b$ işleminin sonucu hesaplanarak verinin sınıfına karar verilir. $y \leq 0$ durumu verinin birinci sınıfa, $y > 0$ durumu ise verinin ikinci sınıfa ait olduğunu temsil etmektedir.



Şekil 3. Destek Vektör Makineleri Hiper Düzlem Seçimi
(Determining Hyper Plane in Support Vector Machines)

DVM diğer makine öğrenmesi tekniklerinden farklı olarak verileri sadece iki sınıfa ayırma işlemi yapmaktadır. Üç ya da daha fazla sınıfa sahip veri setlerinde, bire karşı bir (OVO, one versus one), ya da bire karşı hepsi (OVA,one versus all) ayrıştırma yöntemleri kullanılarak çok sınıflı verilerin sınıflandırılma işlemi DVM ile yapılabilir. Çalışmamızda OVA ayrıştırma tekniği kullanılmıştır. İlgili yöntemde farklı sınıf sayısı kadar ikili sınıflayıcılar oluşturulur. Veriler, ilgilenilen sınıf etiketi 1, diğer tüm sınıftaki veriler -1 olacak şekilde etiketlenerek sınıflandırma işlemi gerçekleştirilir. Bu işlem her bir sınıf için oluşturan modeller için ayrı ayrı tekrarlanır. Sınıfllayıcı modellerden elde edilen sonuçlar birleştirilerek gerçek sınıf bilgisi elde edilir.

2.2.2. Yapay Sinir Ağları (Artificial Neural Networks)

İnsan sinir hücreleri dikkate alınarak tasarlanmış Yapay Sinir Ağları (YSA), denetimli ya da denetimsiz birçok türevi olan güçlü bir makine öğrenmesi tekniğidir. En küçük YSA birimine nöron adı verilir. Nöronlar birleşerek katmanları, kamanlar birleşerek modeli meydana getirmektedir. Bir YSA modeli, girdi katmanı ve çıktı katmanı olmak üzere en az iki katmandan meydana gelmelidir. Girdi katmanında eğitim için kullanılacak veri setinin öznitelik sayısı kadar, çıktı katmanında ise sınıf sayısı kadar nöron bulunmalıdır. Oluşturulan bu iki katmanlı YSA modeli doğrusal olarak ayrışabilen veri setlerinde başarılı sonuçlar verse de, doğrusal olarak ayrışamayan veri setlerini çözmede başarılı olamamıştır. Bu problemi çözmek için girdi katmanı ile çıktı katmanı arasına nöron sayısı veri setinden bağımsız ve değişkenlik gösterebilen bir ya da daha fazla gizli katman eklenmektedir. YSA modelinde bir katmanda bulunan her bir nöron, bir sonraki katmanda bulunan tüm nöronlara bağlı ise bu model Tam Bağlı YSA olarak adlandırılmaktadır. Sınıflama için en sık kullanılan Çok Katmanlı YSA (ÇKYSA) modeli bu üç katman çeşidini de bulunduran tam bağlı bir modeldir. ÇKYSA'da eğitim, oto kodlayıcı modeline benzer bir şekilde eşitlik 5 ve 6 kullanılarak, ileri doğru ve geri besleme olmak üzere iki aşamada ağırlıklar belirlenerek yapılır. Son aşamada ise belirlenen ağırlıklar kullanılarak yeni veriler için sınıf tahmini yapılır ve model sonlandırılır.

3. UYGULAMALAR VE BULGULAR (APPLICATIONS AND RESULTS)

Çalışmada film yorumları içeren 1000 pozitif ve 1000 negatif olmak üzere toplam 2000 adet örnek içeren IMDB veri seti kullanılmıştır [25]. Veri setinden gerekli öznitelikler çıkarılarak sınıflandırmaya hazır hale getirmek amacıyla çeşitli veri ön işleme adımları uygulanmıştır. Metinler kelimelere ayrılmadan önce özel işaretler ve noktalama işaretleri kaldırılmış, tüm metin küçük harflere çevrilmiştir. Daha sonra her bir kelimenin kökleri NLTK kütüphanesi yardımıyla bulunmuş, edat, bağlaç ve zamirlerden oluşan durak kelimeleri ile uzunluğu 3 harften küçük kelimeler silinmiştir. Terim olarak da isimlendirilen birbirinden farklı kelimeler öznitelik olarak ele alınmış, ardından terim frekansları (TF) ve ters doküman Frekansları (IDF) yardımıyla vektör uzay modelleri oluşturulmuştur. Öznitelik olarak TF-IDF değeri en yüksek olan 1000 terim ele alınmıştır. Bu veri seti DVM ve YSA sınıflama algoritmaları kullanılarak ilk önce orijinal boyutta eğitilmiş daha sonra TBA, çTBA, TDA, FA, Oto Kodlayıcı ve Denoising Oto Kodlayıcı olmak üzere 6 farklı boyut düşürme yöntemi uygulanarak yeni veri setleri elde edilmiştir. Son olarak da bu veri setleri üzerinde eğitimler gerçekleştirilmiştir.

Tüm veri setlerinde, veri setinin %75'lik kısmı eğitim için, %25'lik kısmı ise test için rastgele ayrılmıştır. Oto kodlayıcı modelleri dışında her bir boyut düşürme yöntemi için 40. boyuttan başlanarak her adımda 10'ar artacak şekilde 530. boyuta kadar 530 farklı boyutta elde edilen veri setleri DVM ve YSA ile eğitilmiştir. Oto kodlayıcı modellerinde ise boyut ilk olarak tek kademede düşürülmüş daha sonra derin öğrenme tabanlı olacak şekilde iki oto kodlayıcı kullanılarak farklı bir model oluşturulmuştur. Tek kademede boyut düşürme işlemi diğer yöntemlerde olduğu gibi 530 ile 40 boyut arasında boyutlar değiştirilerek yapılmıştır. İki kademeli boyut düşürme işlemi ise ilk kademede 800 ile 600, ikinci kademede 530 boyut ile 40 boyut arasında olacak şekilde eğitilmiştir. Bahsedilen her bir yöntem için test verilerinde en yüksek başarı oranını veren boyut tespit edilmiştir.

Sonuçlar incelendiğinde YSA sınıflama algoritması için en iyi başarı oranları; Oto Kodlayıcı için 400 boyuta düşürülen iki kademeli derin mimaride, Denoising Oto Kodlayıcı için 390 boyuta düşürülen iki kademeli derin mimaride, FA için 250. boyutta, TBA için 210. boyutta, çTBA için 210. boyutta, TDA için 110. boyutta alınmıştır. DVM sınıflama algoritması için en iyi başarı oranları; Oto Kodlayıcı için 360 boyuta düşürülen tek kademeli yapıda, Denoising Oto Kodlayıcı için 300 boyuta düşürülen tek kademeli yapıda, FA için 320. boyutta, TBA için 320. boyutta, çTBA için 460. boyutta, TDA için 210. boyutta

alınmıştır. Tespit edilen en iyi modeller için ve orijinal boyutta eğitilen model için test verisine ait doğruluk (Accuracy - Acc), hassasiyet (Sensitivity - Sens), özgüllük (Specificity - Spec), ve f-ölçüt (F) değerlerleri Tablo 1'de gösterilmiştir. Bu tabloda yöntem sütundaki her bir satır geliştirilmiş modelleri temsil etmektedir. M0; DVM'nin öznitelik düşürme yapılmadan kullanıldığı, M1; TDA ile DVM yöntemlerinin birlikte kullanıldığı, M2; TBA ile DVM yöntemlerinin birlikte kullanıldığı, M3; çTBA ile DVM yöntemlerinin birlikte kullanıldığı, M4; FA ile DVM yöntemlerinin birlikte kullanıldığı, M5; Oto Kodlayıcı ve DVM yöntemlerinin birlikte kullanıldığı, M6; Denoising Oto kodlayıcı ile DVM yöntemlerinin birlikte kullanıldığı, M7; YSA'nın öznitelik düşürme yapılmadan kullanıldığı, M8; TDA ile YSA yöntemlerinin birlikte kullanıldığı, M9; TBA ile YSA yöntemlerinin birlikte kullanıldığı, M10; çTBA ile YSA yöntemlerinin birlikte kullanıldığı, M11; FA ile YSA yöntemlerinin birlikte kullanıldığı, M12; Oto Kodlayıcı ile YSA yöntemlerinin birlikte kullanıldığı, M13 ise Denoising Oto Kodlayıcı ile YSA yöntemlerinin birlikte kullanıldığı modelleri belirtmektedir.

Tablo 1. Analiz Sonuçları
(Analysis Results)

Yöntem	Boyut	Acc	Sens	Spec	F
M0	1000	0.7800	0.7591	0.8053	0.7909
M1	210	0.8040	0.8175	0.7876	0.8205
M2	320	0.8140	0.8066	0.8230	0.8262
M3	460	0.7740	0.7847	0.7611	0.7919
M4	320	0.8080	0.8139	0.8009	0.8229
M5	360	0.8080	0.8723	0.7301	0.8328
M6	300	0.8020	0.8212	0.7788	0.8197
M7	1000	0.7860	0.7701	0.8053	0.7977
M8	110	0.7900	0.7774	0.8053	0.8023
M9	210	0.8120	0.7956	0.8319	0.8226
M10	210	0.7920	0.7847	0.8009	0.8052
M11	250	0.7900	0.7810	0.8009	0.8030
M12	400	0.8160	0.8358	0.7920	0.8327
M13	390	0.8160	0.8321	0.7965	0.8321

Performans ölçütlerinden görüleceği üzere boyut düşürme tekniklerinin, istisnai bir durum olan çTBA ile DVM yöntemlerini birlikte kullanan model haricinde daha iyi bir başarı oranı elde ettiği gözlemlenmiştir. Boyut düşürme teknikleri kendi aralarında kıyaslandığında; DVM'nin sınıflama algoritması olarak kullanıldığı modeller için, önerilen iki farklı derin öğrenme tabanlı oto kodlayıcı modeli başarı oranı olarak diğer boyut düşürme tekniklerine yakın sonuçlar almışlardır. YSA'nın sınıflama algoritması olarak kullanıldığı modellerde ise oto kodlayıcılar diğer yöntemlerden daha iyi başarı elde etmişlerdir. YSA modellerinden oto kodlayıcı tekniklerinin daha iyi sonuçlar almasının nedeni, bu iki yöntemin yapı olarak birbirine benzemesi ve daha iyi uyum sağlaması olarak değerlendirilebilir. Her iki sınıflama algoritması içinde, doğru tahmin edilen pozitif duyguların oranını veren hassasiyet skorunda, önerilen oto kodlayıcı modellerinin kayda değer oranda daha iyi sonuçlar verdiği gözlemlenmektedir. Oto kodlayıcı modelleri kendi aralarında kıyaslandığında; YSA modeli için çok benzer sonuçlar vermekle birlikte, DVM için oto kodlayıcı modeli, denoising oto kodlayıcı modeline göre daha iyi sonuçlar vermiştir. Bu fark, veri setinde bol sıfır olması ve eklenen gürültünün veri setinde bazı bozulmalara neden olması olarak değerlendirilmiştir.

4. SONUÇ (CONCLUSION)

Bu çalışmada metin madenciliğinin bir alt dalı olan duygu analizi için boyut düşürme teknikleri önerilmiştir. Metin madenciliğinde öznitelik çıkarma tekniklerinin yapısı gereği, veri setlerinin karmaşık bir hal alması analizi zorlaştırmaktadır. Analizi daha kolay bir hale getirmek için IMDB film yorumlarından elde edilen veri seti, DVM ve YSA sınıflama algoritmaları kullanılarak ilk adımda orijinal boyutta eğitilmiştir. İkinci adımda veri seti belirlenen boyut düşürme teknikleri ile ön işlemden geçirilerek sınıflamaya en uygun boyut tespit edilmiştir. Son adımda ise tespit edilen boyutlardaki veriler ile sistem yeniden eğitilerek sonuçlar değerlendirilmiştir. Elde edilen sonuçlar doğrultusunda boyut düşürme tekniklerinin beklendiği gibi başarı oranını artırdığı gözlemlenmiştir. Ancak çTBA boyut düşürme tekniğinin bu veri seti üzerinde başarı oranında kayda değer bir iyileştirme yapmadığı, önerilen derin öğrenme tabanlı oto kodlayıcı modellerin ise diğer yöntemlere kıyasla benzer ya da daha iyi sonuçlar elde ettiği, bu yöntemlerin bazı durumlarda kayda değer iyileştirmeler yaptığı sonuçlara bakılarak anlaşılmıştır.

KAYNAKLAR (REFERENCES)

[1] J. Li and M. Sun, "Experimental Study on Sentiment Classification of Chinese Review using Machine Learning Techniques," **International Conference on Natural Language Processing and Knowledge Engineering**, 2007, pp. 393–400.

[2] G. Alec, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision." 2009.

[3] K. Mouthami, K. N. Devi, and V. M. Bhaskaran, "Sentiment analysis and classification based on textual reviews," **International Conference on Information Communication and Embedded Systems (ICICES)**, 2013, pp. 271–276.

[4] V. K. Singh, R. Piryani, A. Uddin, P. Waila, and Marisha, "Sentiment Analysis of Textual Reviews," **The 5 th International Conference on Knowledge and Smart Technology**, 2013.

[5] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," **Seventh International Conference on Contemporary Computing (IC3)**, 2014, pp. 437–442.

[6] H. Nizam and S. S. Akın, "Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması," **XIX. Türkiye'de İnternet Konferansı**, 2014.

[7] Ö. Çoban, B. Özyer, and G. T. Özyer, "Sentiment analysis for Turkish Twitter feeds," **23nd Signal Processing and Communications Applications Conference (SIU)**, 2015, pp. 2388–2391.

[8] J. Kranjc, J. Smailović, V. Podpečan, M. Grčar, M. Žnidaršič, and N. Lavrač, "Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform," *Inf. Process. Manag.*, vol. 51, no. 2, pp. 187–203, Mar. 2015.

[9] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst. Appl.*, vol. 57, pp. 117–126, Eylül 2016.

[10] V. Rohini, M. Thomas, and C. A. Latha, "Domain based sentiment analysis in regional Language-Kannada using machine learning algorithm," **IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)**, 2016, pp. 503–507.

[11] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Inf. Sci.*, vol. 181, no. 6, pp. 1138–1152, Mar. 2011.

[12] M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," **Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)**, 2013, pp. 1–5.

[13] E. Fersini, E. Messina, and F. A. Pozzi, "Sentiment analysis: Bayesian Ensemble Learning," *Decis. Support Syst.*, vol. 68, pp. 26–38, Aralık 2014.

[14] N. F. F. da Silva, E. R. Hruschka, and E. R. Hruschka Jr., "Tweet sentiment analysis with classifier ensembles," *Decis. Support Syst.*, vol. 66, pp. 170–179, Ekim 2014.

[15] C. Catal and M. Nangir, "A sentiment classification model based on multiple classifiers," *Appl. Soft Comput.*, vol. 50, pp. 135–141, Ocak 2017.

[16] S. Tan and J. Zhang, "An empirical study of sentiment analysis for chinese documents," *Expert Syst. Appl.*, vol. 34, no. 4, pp. 2622–2629, May 2008.

[17] A. GO, L. Huang, and R. Bhayani, "Twitter Sentiment Analysis." 2009.

[18] M. Meral and B. Diri, "Sentiment analysis on Twitter," **22nd Signal Processing and Communications Applications Conference (SIU)**, 2014, pp. 690–693.

[19] G. Vinodhini and R. M. Chandrasekaran, "Effect of Feature Reduction in Sentiment analysis of online reviews," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 2, no. 6, 2013.

[20] A. Yousefpour and H. N. Hamed, "A Novel Feature Reduction Method in Sentiment Analysis," *Int. J. Innov. Comput.*, 2014.

[21] K. Kim and J. Lee, "Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction," *Pattern Recognit.*, vol. 47, no. 2, pp. 758–768, ubat 2014.

[22] L. B. Shyamasundar and P. J. Rani, "Twitter sentiment analysis with different feature extractors and dimensionality reduction using supervised learning algorithms," **IEEE Annual India Conference (INDICON)**, 2016, pp. 1–6.

[23] P. Baldi, "Autoencoders, Unsupervised Learning, and Deep Architectures," **Workshop on Unsupervised and Transfer Learning**, 2012.

[24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1," **D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds. Cambridge**, MA, USA: MIT Press, 1986, pp. 318–362.

[25] Movie Review Data: https://www.cs.cornell.edu/people/pabo/movie-review-data/, 01.03.2017.