Journal of the Faculty of Engineering and Architecture of Gazi University 36:2 (2021) 701-714

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Document Sentiment classification using hybrid wavelet methodologies

İlknur Dönmez[1]*, Zafer Aslan[2]

[1]Department of Computer Engineering, İstanbul Arel University, 34537, İstanbul, Turkey
[2]Department of Computer Engineering, İstanbul Aydın University, 34295, İstanbul, Turkey

Highlights:

- Haar transformation in text analysis
- Using wavelet in sentiment analysis
- Text visualization using wavelet

Keywords:

- Wavelet
- Sentiment Analysis
- Text mining
- Text visualization
- Haar Transformation

Article Info:

Research Article
Received: 09.03.2020
Accepted: 11.10.2020

DOI:

10.17341/gazimmfd.701313

Correspondence:

Author: İlknur Dönmez
e-mail: buyukkuscu@itu.edu.tr
phone: +90 505 392 9366

Graphical/Tabular Abstract

Sentiment and semantic analysis of a text are very important issues of today because of increasing text data. Our study proposes a new method to reveal the hypernym relations (generic term) of the words in the text and to enhance the accuracy of sentiment classification result of the texts. We used wavelet transform method that has been rarely used in text analysis.

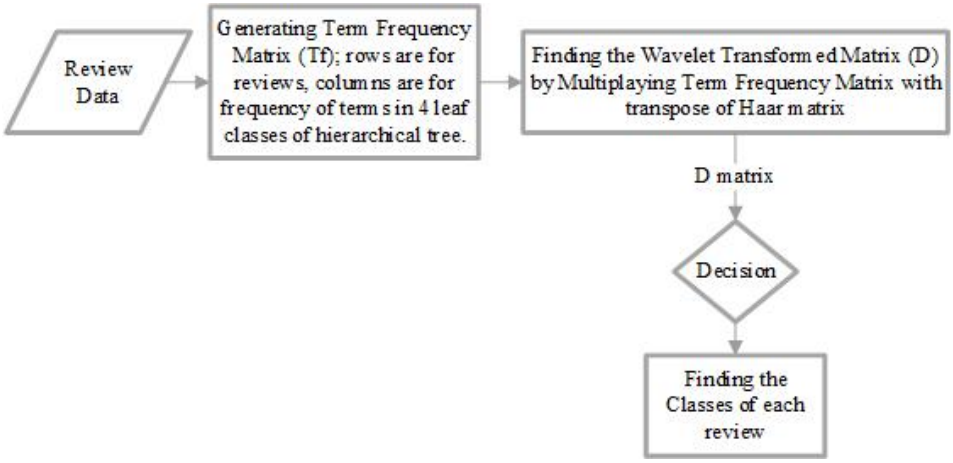


Figure A. General Flow Diagram of Wavelet Sentiment Classification

Purpose: The aim of the study is finding new approaches that uses wavelet transformation technique for text semantic representation and text sentiment analysis. We investigate the contribution of wavelet on the sentiment analysis classification problem. We used classical algorithms and hybrid wavelet algorithm for sentiment analysis problem.

Theory and Methods:
The wavelet transformation, which is generally utilized for signal processing, is used for text semantic representation and text sentiment analysis. The frequency vectors that include term frequency of special words in the reviews are transformed to D matrix using a discrete wavelet transformation. The flow of the wavelet sentiment classification is seen on Figure A. We used 5 polarity classes, which are “high positive”, “low positive”, “neutral”, “high negative”, and “low negative” in our classification problem.

Results:
Even wavelet is generally used for signal processing, it is seen that it is also useful for text semantic representation and sentiment analysis. In our study, to use wavelet transformation with classical classification algorithm provided % 3-5 increases in accuracy.

Conclusion:
As a result, wavelet transform can be used in the semantic representation of the text. When the wavelet method is added to the classical sentiment classification problem, it helps to improve accuracy. It is expected that our method will contribute to text analysis problems and applications such as decision making and text applications containing hierarchical data.

Journal of the Faculty of Engineering and Architecture of Gazi University 36:2 (2021) 701-714

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University
Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Metin Duygu sınıflandırılmasında hibrit wavelet yönteminin kullanımı

İlknur Dönmez[1]*, Zafer Aslan[2]
[1]İstanbul Arel Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, 34537, Tepekent, İstanbul, Türkiye
[2]İstanbul Aydın Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 34295, Florya İstanbul, Türkiye

ÖNEÇIKANLAR

- Metin analizinde Haar dönüşümü
- Duygu analizinde dalgacık kullanma
- Dalgacık kullanarak metin görselleştirme

Makale Bilgileri
Araştırma Makalesi
Geliş: 09.03.2020
Kabul: 11.10.2020

DOI:
10.17341/gazimmfd.701313

Anahtar Kelimeler:
Dalgacık
duygu analizi,
metin madenciliği,
metin görselleştirme,
haar dönüşümü

ÖZET

Verilerin her geçen gün arttığı günümüzde herhangi bir metnin anlamsal ve duygusal çözümlemesi ihtiyaç duyulan konulardan biridir. Çalışmamız metinlerin sınıflandırılmasında kullanılabilecek üst anlam ilişkilerini çıkarmak ve metinlerin duygu sınıflandırmasını yapmak için yeni bir yöntem önermektedir. Bu yöntem daha önce metin analizinde çok az kullanılmış Wavelet (Dalgacık Dönüşüm) yöntemidir. Amacımız bu yöntemin duygu analizi sınıflandırma probleminde nasıl katkı sağladığını göstermektir. Çalışmada klasik sınıflandırma algoritmalarının yanında hibrit wavelet yöntemi kullanılmıştır. Klasik sınıflandırma algoritmalarına wavelet eklendiğinde doğrulukların yaklaşık olarak %5 arttığı gözlemlenmiştir.

Document Sentiment classification using hybrid wavelet methodologies

HIGHLIGHTS

- Haar transformation in text analysis
- Using wavelet in sentiment analysis
- Text visualization using wavelet

Article Info
Research Article
Received: 09.03.2020
Accepted: 11.10.2020

DOI:
10.17341/gazimmfd.701313

Keywords:
Wavelet,
sentiment analysis,
text mining,
text visualization,
haar transformation

ABSTRACT

Sentiment and semantic analysis of a text are very important issues of today because of increasing text data. Our study proposes a new method to reveal the hypernym relations (generic term) of the words in the text and to enhance the accuracy results of sentiment classification of the texts. We used wavelet transform method that has been rarely used in text analysis. In our study, the aim is to show how this method contributes the sentiment analysis classification problem. We used classical algorithms and hybrit wavelet algorithm for sentiment analysis problem. It has been observed that when wavelets are applied to classical classification algorithms, the accuracies increased approximately 5%.

*Sorumlu Yazar/Yazarlar / *Corresponding Author/Authors: *ilknurdonmez@arel.edu.tr, zaferaslan@aydin.edu.tr / Tel: +90 505 392 9366

1. GİRİŞ (INTRODUCTION)

Dalgacık dönüşümü (Wavelet), verilen bir sinyalin zaman-frekans gösterimini sağlar. Dalgacık dönüşümünün, Fourier dönüşümünün kullanılıp yetersiz kaldığı veri analizi durumlarında, başarılı sonuçlar verdiği görülmüştür (Akansu, 1995 [1]; Chan, 1996[2]; Meyer, 1993[3]; Strang, 1996 [4]). Özellikle değişimi küçük veya düzensiz ayrıntılara sahip işaretler ve görüntüler dalgacıklar dönüşümü ile Fourier'e göre genellikle daha iyi analiz edilebilir (Meyer, 1993 [3]). Dalgacık dönüşümü, ses ve işitsel işaret işleme, görüntü ve video işleme, haberleşme, jeofizik, ekonomi ve tıp gibi özellikle bir boyutlu ve iki boyutlu işaret işleme uygulamalarında yoğun olarak kullanılmaktadır. Dalgacık dönüşüm konusunda teorinin ayrıntılarını içeren pek çok çalışma mevcuttur [5-9].

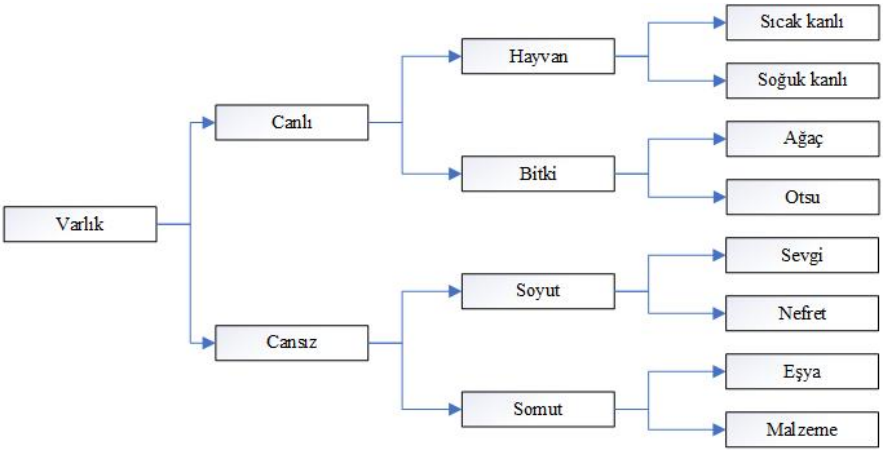
Dalgacık dönüşümü 1990'lardan beri farklı veri türleri üzerinde denenmiş olmasına karşın metin üzerine dalgacık dönüşümü nispeten yeni ve daha az çalışılmış bir konudur. 2002 yılında veri madenciliği alanında dalgacık uygulamaları üzerine bir araştırma çalışması mevcuttur [10]. 2002 yılında bir başka çalışmada Agarwal DNA karakter dizileri için dalgacık dönüşümü kullanmıştır [11]. Bu çalışmada Dalgacık tekniğinin, farklı ayrıntı seviyelerinde eğilimleri (trendleri) yakalayabilen ve böylece yeni temsil ile sınıflandırma görevine yardımcı olan hiyerarşik bir ayrışma oluşturduğunu iddia edilmektedir. 2007'de Chao Xu ve Yi-Ming Zhou wavelet dönüşümünü ilk kez doküman sınıflandırma probleminde kullanmaktadır [12]. Bu çalışmada dokümanda bulunan terimlerin üst sınıf temsiline değinilmiştir. 2008'de Geraldo Xexeo dalgacık dönüşümünü metinlerin demetlenmesi ile ilgili uygulamasında kullanılmıştır [13]. Dokümanlar için bulunan terim sıklık vektörleri bir birine yakınlığına göre sıralanıp elde edilen matrise dalgacık dönüşümü uygulanmıştır. Elde edilen dönüştürülmüş doküman vektörleri benzerlik ölçütleri kullanılarak demetlenir. 2015 yılında yapılan bir başka çalışmada dalgacık dönüşümü kısa metinler üzerinde kullanılmıştır. Metnin vektör temsili üzerine wavelet dönüşümü uygulanarak, en önemli (ağırlığı en fazla olan) özellikler araştırılmıştır [14]. Bu araştırma çalışmasında, dalgacık dönüşümünün hiyerarşik anlamsal yapıdaki kelime ilişkilerini nasıl temsil edeceği açıklanmış, şu ana kadar yapılan çalışmalara ek olarak, wavelet dokümanın içinde geçen duyguların görselleştirmesi amacıyla kullanılmıştır. Bu makalede, doküman duygu analizi ve görselleştirmesi için dalgacık dönüşümünü kullanan yeni bir yöntem önerilmektedir.

Çalışmamızın ikinci bölümü wavelet dönüşümünün duygu analizi amacıyla kullanımı ile ilgilidir. Üçüncü kısımda duygu analizi sınıflandırması problemleri için kullanılan algoritmalar ve önerdiğimiz wavelet modeli açıklanmaktadır. Dördüncü bölümde wavelet ile duygu analizi görselleştirmesi ve wavelet kullanımının sınıflandırma doğruluklarını nasıl etkilediğine dair sonuçlar yer almaktadır. Makalede, ayrıca tartışma ve sonuç kısımları bulunmaktadır.

2. WAVELET DÖKÜMAN ANALİZ MODELİ (WAVELET DOCUMENT ANALYSIS MODEL)

Nasıl ki dalgacık dönüşümü (Wavelet), bir sinyalin farklı frekanslarda detaylandırıp ifade etmek için kullanılabiliyorsa; bu bir dokümanın farklı sınıflara ait terimlerle temsil edilmesine benzetilebilir. Örneğin "sevgi" kelimesi söylendiğinde; bunun "soyut" olduğu "cansız" olduğu bir "varlık" olduğu bu kelimenin içinde gizlidir. Hiyerarşik yapının en tepesinde düşük frekanslı sinyaller; en altında ise yüksek frekanslı sinyaller vardır. Kelime üst anlam ilişkileri ağacında düğümlerde detaylı ve kısıtlı bilgiler (ağaç); üste çıktıkça daha az detaylı ve kapsayıcı bilgiler (canlı veya varlık) bulunmaktadır.

Çalışmamızda bir dokümanın her bir paragrafı için paragrafta geçen kelimelerin; hiyerarşik ağaç yapısının yaprak düğümlerindeki hangi sınıfa ait olduğu belirlenir. Örneğin Şekil 1'deki gibi bir hiyerarşik ağaç yapısı için; dökümanda "kuş, kedi, arslan" geçiyorsa bu kelimeler "sıcak kanlı" sınıfındadır. Yine "aşk, hoşlanmak, ilgi" geçiyorsa sevgi sınıfının kelimeleridir. Çalışmamızda tüm dökümanda



Şekil 1. Hiyerarşik varlıkbilim ağacı (Hierarchical ontological tree)

yaprak sınıflardaki kelime sayıları bir vektör olarak ifade edilmekte bu da terim frekans vektörünü oluşturmaktadır. Bu vektör ana dalgacığı oluşturmaktadır. Ana dalgacık ayrık dalgacık dönüşüm matrisiyle çarpılarak aynı uzunluklu bir başka vektör elde edilir. Dönüşüme uğramış bu yeni vektörün katsayıları Şekil 1'deki hiyerarşik ağaçtaki her bir ikili ağacın baskınlığı hakkında bilgi vermektedir. Yani ağacın ikili dallarında benzer terimlerin zıt halleri mevcuttur. Şekil 1'deki ikili alt ağaçlarda üstteki terim pozitif işaretli, alttaki terim negatif işaretli gibi davranır. Hangisinin ağırlığı baskınsa o terim ön plana çıkar. Eğer iki tarafın ağırlıkları birbirine eşitse; özellik 0'lanır.

Tablo 1'de bir metnin Şekil 1'deki hiyerarşik ağacın en alt sınıflarına ait olan terimlerinin sayısı gösterilmektedir. Bu doküman T_f=[2,0,0,1,1,1,0,0] ile gösterilmektedir.

Tablo 1. Bir metnin terim frekans vektörünün çıkarılması
(To generate term frequency vector for a document)

Yaprak Düğüm (Uç Sınıflar)	Dökümandaki Sınıfa Ait Terimler	T_f
Sıcakkanlı	Aslan, zürafa	2
Soğukkanlı	0	0
Ağaç	0	0
Yeşillik	Kaktüs	1
Sevgi	Sevmek	1
Nefret	Kin	1
Eşya	0	0
Malzeme	0	0

Terimleri sınıflarda aranıp bulunan T_f vektörü Eş. 1'de gösterilmektedir. 8x8'lik Haar dönüşüm matrisi Eş. 2'de gösterilmektedir. Eş. 3'de gösterildiği üzere terim frekans vektörü T_f ayrık dalgacık dönüşüm matrisinin devriği (transpozu) ${H_8}^T$ ile çarpılarak Eş. 4'de gösterilen D dönüşüm vektörü elde edilir.

$$T_f = [2,0,0,1,1,1,0,0] \tag{1}$$

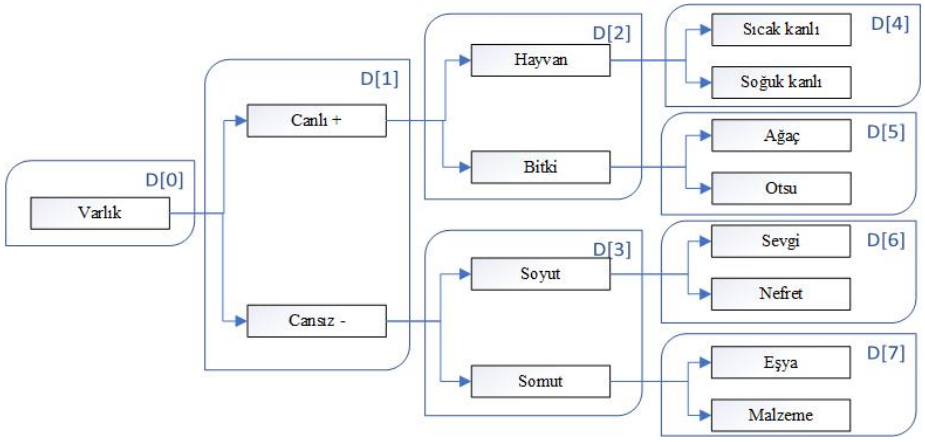
$$H_8 = \begin{vmatrix} \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} \\ \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} \\ \sqrt{\frac{1}{4}} & \sqrt{\frac{1}{4}} & -\sqrt{\frac{1}{4}} & -\sqrt{\frac{1}{4}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{\frac{1}{4}} & \sqrt{\frac{1}{4}} & -\sqrt{\frac{1}{4}} & -\sqrt{\frac{1}{4}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{vmatrix} \tag{2}$$

$$D = T_f x {H_8}^T \tag{3}$$

$$D = \left| 5\sqrt{\frac{1}{8}} \quad \sqrt{\frac{1}{8}} \quad \sqrt{\frac{1}{4}} \quad 2\sqrt{\frac{1}{4}} \quad 2\sqrt{\frac{1}{2}} \quad -\sqrt{\frac{1}{2}} \quad 0 \quad 0 \right| \tag{4}$$

Elde edilen D vektörü özel bir vektör olup; hiyerarşik ağacın her bir elemanı ile ilgili bilgi içerir. Şekil 2'de gösterildiği üzere, D vektörünün 0 indeksli ilk elemanında kullanılan $5\sqrt{\frac{1}{8}}$ çarpanı varlık sayısının 5 olduğunu ve bu grubun genel çarpanının $\sqrt{\frac{1}{8}}$ olduğunu ifade etmektedir (aslan, zürafa, kaktüs, sevmek, kin). D vektörünün bir indeksli ikinci terimi; + işaretli canlı sınıfı ve – işaretli cansız sınıfından hangisinin baskın olduğunu göstermektedir. "Canlı" sınıfına ait 3 terim (aslan, zürafa ve kaktüs) varken, "Cansız" sınıfına ait 2 terim (sevmek, kin) vardır (Tablo 1'den kontrol edilebilir). Canlı sınıfı cansız sınıfına göre 1 birim baskındır. Pozitif işaretli canlı sınıfı $\sqrt{\frac{1}{8}}$ katsayısı ile gösterilmektedir. D vektörünün iki indeksli terimi hayvan bitki sınıflarının birbiriyle üstünlüğü ile ilgilidir. "Hayvan"la ilgili 2 tane "Bitki"yle ilgili 1 tane terim olduğundan pozitif işaretli hayvan terimi $\sqrt{\frac{1}{4}}$ katsayısı ile gösterilmektedir.

D'nin üç indeksli elemanı soyut ve fiziksel sınıflarından hangisinin baskın olduğuyla ilgilidir. 2 soyut (sevmek, kin) ve 0 fiziksel varlık olduğundan soyut terimler baskındır ve



Şekil 2. D matrisindeki ağırlıkların üst sınıflarla bağlantısı (The relation between the weights of D matrix and upper classes)

çarpanı $2\sqrt{\frac{1}{4}}$'dür. D'nin beş indeksli elemanı ağaç ve yeşillikten hangisinin baskın olduğunu söyler. "Ağaç"la ilgili hiç terim olmamasına rağmen "Yeşillik"le ilgili 1 terim olduğundan katsayı -$\sqrt{\frac{1}{2}}$ olur. D'nin altı indeksli elemanı için sevgi sınıfında 1 nefret sınıfında 1 terim bulunduğu için katsayı 0 olmuştur. Son eleman içinse eşya ve malzemeye ait terim olmadığından katsayı 0'dır. Bu kadar uyumlu sonuçların çıkması terim frekans vektörünü çarptığımız Haar dönüşüm matrisinin iç yapısından yani katsayılarından kaynaklanmaktadır.

Tablo 2 ve Tablo 3'de görüldüğü üzere terim frekans matrisi Haar dönüşüm matrisinin devriği ile çarpıldığında oluşan D vektörünün elemanları dokümanda görülen terimlerin üst sınıfları hakkında bilgi içermektedir. D dönüşüm vektörünün elemanının işareti, ikili sınıflardan hangisinin dikkate alınacağını belirler.

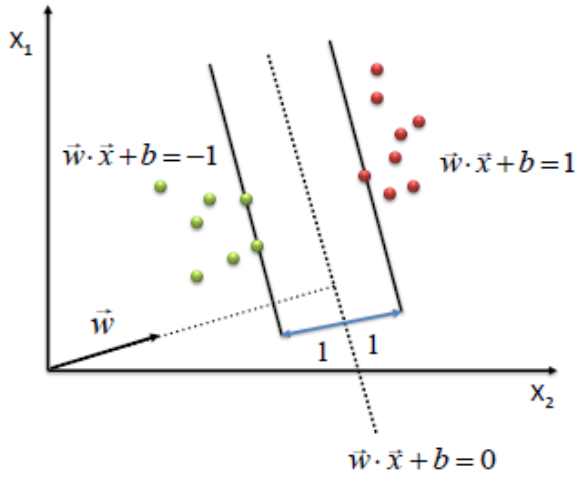
Yeni oluşturulan D wavelet dönüşüm vektörü özel bir vektör olup, üst sınıfların ağırlıklarını içermektedir. Bu şekilde, metinde geçen çok detaylı olduğu için doküman sınıflarında bulunamayan sözcükler, üst sınıflarıyla bir başka ifade ile kapsayıcı terimi ile (hypernym) ifade edildiğinde bulunabilir. Örneğin, "Armonica" kelimesinin sık geçtiği bir metnin sınıfını bulmak istiyor olalım. Sınıflarımız spor, haber ve müzik olsun. Armonika anahtar kelimesi çok özel bir enstrümana ait olup, müzikle ilgili bir metinde dahi bulunmayabilir; ama üst sınıfın müzik olduğu bilgisi de mevcut olsaydı, bu kelimenin sık geçtiği bir metni müzikle ilgilidir şeklinde etiketlemek kolay olurdu. Yani sınıflandırma algoritmalarında terim frekans vektörlerinin yanında; terimlerin üst sınıflarını da kullanmanın (D vektörünü de kullanmanın) sınıflandırmanın doğruluğunu artıracağı söylenebilir. Kullandığımız wavelet dönüşüm vektörünün elemanları, terimlerin üst anlamlarını temsil eden ağırlıkları oluşturur. Buradaki önemli konulardan biride hiyerarşik ağaçta aşağıdan (yaprak düğümler) yukarıya doğru (kök) ağırlıklar azalmaktadır. Bunun nedeni özel bir kelimeden onun üst anlamına geçtiğimizde kapsam artarken, detayı kaçırmaya başlarız yani üst terimlerin verdiği bilgi azalır.

3. METİN DUYGU ANALİZİ
(DOCUMENT SENTIMENT ANALYSIS)

Duygu analizi, insanların fikirleri, tutumları hakkında var olan bir metinden yararlanılarak metnin öz duygusunun çıkarılmasına olanak vermektedir [15]. Verinin artmasıyla birlikte duygu analizi son yıllarca çokça çalışılan konulardan biridir [16-20]. Çalışmamızda duygu analizi uygulaması için destek vektör makinası, çok katmanlı yapay sinir ağları, çok değişkenli yapay sinir ağları, wavelet yöntemi kullanılmıştır. Bir sonraki aşamada önerdiğimiz wavelet yöntemi diğer algoritmalarla birleştirildiğinde doğruluklarda artış sağlanıp sağlanmadığı test edilmiştir.

3.1. Destek vektör makinesi ile Duygu Analizi
(Sentiment Analysis using Suport Vector Machine)

Destek vektör makinesi iki boyutlu uzayda doğrusal, üç boyutlu uzayda düzlemsel ve çok boyutlu uzayda hiper-düzlem şeklindeki ayırma mekanizmaları kullanarak veriyi iki ya da daha çok sınıfa ayırmayı sağlayan bir sınıflandırma yöntemidir. Bu algoritma Lineer olarak ayrıştırılabilen verilerin sınıflarının belirlenmesinde kullanılabildiği gibi, kernel fonksiyonları ve iz düşüm yöntemleri sayesinde doğrusal olarak ayrıştırılamayan verilerin sınıflandırılmasında da başarıyla kullanılmaktadır. Eğitim için kullanılacak N tane verinin Q={X_i,y_i}, i=1,2,N olduğu varsayılırsa; X_i özellik vektörünü, y_i={-1,1} ise sınıf değerlerini gösterir. Doğrusal ayrılabilen veriler bir ayırıcı düzlem ile ayrılabilirler. Veri setini sınıflara ayırabilecek sonsuz sayıda çoklu düzlem çizilebilmesine karşın, Destek Vektör Makinesinde amaç, sınıflama hatasının en küçük olacağı hiper-düzlemi seçmektir.



Şekil 3. Destek vektör makinesi modeli
(Suport Vector Machine model)

Tablo 2. T_f terim frekans vektörünün içeriği (Contents of T_f vector)

Sınıf: (Terim)	Sıcakkanlı (aslan,zürafa)	Soğukkanlı	Ağaç	Yeşillik (kaktüs)	Sevgi (sevmek)	Nefret (kin)	Eşya	Malzeme
T_f:	2	0	0	1	1	1	0	0

Tablo 3. D wavelet dönüşüm vektörünün içeriği (Contents of D vector)

Üst sınıf:	Varlık	Canlı (+)	Hayvan (+)	Soyut(+)	Sıcakkanlı(+)	Yeşillik (-)	Sevgi/Nefret(0)	Eşya/Malzeme (0)
D:	$5\sqrt{1/8}$	$\sqrt{1/8}$	$\sqrt{1/4}$	$2\sqrt{1/4}$	$2\sqrt{1/2}$	$-\sqrt{1/2}$	0	0

Şekil 3'de görüldüğü üzere $f(x)=w^T x+b\geq 1$ durumu birinci sınıfı ($y_i=1$); ve $f(x)=w^T x+b\leq 1$ durumu ikinci sınıfı ($y_i=-1$) temsil eder. İki sınır arasındaki uzaklık λ-$2/w^2$ formülü ile ifade edilir. Amaç, y_i $(w^T x_i+b)-1\geq 0$, $y_i \in \{-1,1\}$ koşulu sağlanarak λ değerini maksimum yapmaktır. SVM duygu analizi uygulamalarında sıklıkla kullanılmaktadır [21-23]. Duygu analizi probleminde yorum içindeki anlamsal katkı sağlayan kelimeler birer özellik olarak kabul edilir ve her özellik vektörün bir boyutunu oluşturur. Burada anlamsal katkı sağlayan kelimeler cümleler içindeki ekler, zarflar gibi cümle duygusuna katkı sağlamayan kelimeler çıkarılıp, kalan kelimelerin kökü bulunarak elde edilir.

3.2. Çok Katmanlı Yapay Sinir Ağları ile Duygu Analizi
(Sentiment Analysis using Multi Layer Perceptron)

Yapay Sinir Ağları(YSA) insan sinir sistemini taklit ederek öğrenmeyi hedefleyen denetimli bir makine öğrenmesi yöntemidir. Şekil 4'de gösterildiği üzere, Yapay sinir ağı çok katmanlı nöronların bağlanması ile oluşturulur. En sık kullanılan yapay sinir ağı modellerinden biri olan çok katmanlı algılayıcı sinir ağı (MLP) modeli; giriş katmanı, gizli katman ve çıktı katmanı olarak üç katmandan oluşmaktadır. Giriş katmanı verilerin okunduğu katmandır. Bu katmanda her bir nöron bir özelliği temsil ettiği için giriş katmanında özellik sayısı kadar nöron bulunması gerekir. Çıktı katmanı ise sınıfların belirlendiği katmandır ve sayısal değer alabilen tek bir nöron veya lojik özellikte sınıf sayısı kadar nörondan oluşur.

Gizli katman ise verilerin ara işleme maruz kaldığı katmandır. Gizli katman sayısı ve bir gizli katmandaki nöron sayısı tasarımcıya bağlı olmamakla birlikte eğitimin kalitesini etkileyen önemli iki faktördür. MLP modelinde her katmandaki nöron, kendine gelen değerlerin ağırlıklar ile çarpımının toplamına eşittir. Kullanılan eğitim algoritması hatanın karesini en aza indirecek şekilde ağırlıkları güncellemeyi hedefler.

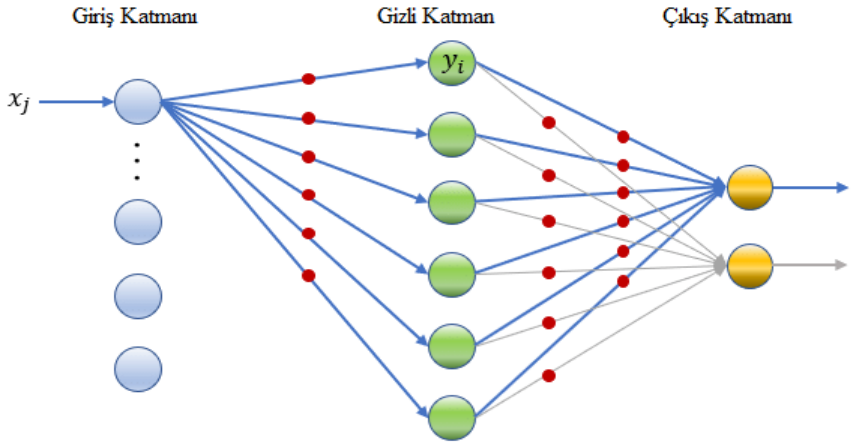
$$y_i = f(\sum_{i=1}^{n} x_j * w_{ji}); \quad (5)$$

$$E(w) = \frac{1}{2}(\sum_{k\ \epsilon\ çıkış} t_k - o_k)^2 \quad (6)$$

Eş. 5, MLP modelinde ileri beslemeyi, Eş. 6 ise geri beslemeyi formüle etmektedir. Eş. 5'te x_j mevcut katmandaki j. nöronun değerini, y_i takip eden katmandaki i. nörona aktarılan değeri, n mevcut katmandaki nöron sayısını, w_{ji} mevcut katmandaki j. nörondan takip eden katmandaki i. nörona giden ağırlığı, f ise aktivasyon fonksiyonunu (örneğin: Gauss, Softmax, Sigmoid) temsil etmektedir. Eş. 6'da k, veri setindeki örnek sayısını, t_k verilerin gerçek sınıfını, o_k ise modelin üretmiş olduğu sınıf değerini temsil etmektedir. MLP duygu analizi uygulamalarında kullanılan bir yöntemdir [24-26]. Dang et al. (2020) tarafından yapılan araştırma çalışmasında, derin öğrenme modellerinin ve sahip olduğu ilgili tekniklerin özü üzerinde durulmakta olup, facebook vb. sosyal ağ verileri için duygu analizi uygulanmış, kelime yerleştirme ve TF-IDF kullanılmıştır.[27]. Duygu analizi uygulamamızda yorumları oluşturan anahtar kelimelerin her birinin kelime vektör değerleri özellik olarak kabul edilmekte ve girişleri oluşturmaktadır.

3.3. Çok değişkenli Doğrusal Regresyon ile Duygu Analizi
(Sentiment Analysis using Multi-Linear Regression)

Regresyon analizi, aralarında sebep-sonuç ilişkisi bulunan iki veya daha fazla değişken arasındaki ilişkiyi belirleyen ve bu ilişkiyi kullanarak o konu ile ilgili tahminler yapan bir analiz yöntemidir. Regresyon yöntemleri duygu analizi uygulamalarında kullanılmaktadır [28-30]. Çalışmamızda duygu analizi için tahmin edilecek 5 değeri sürekli bir aralığa kaydırdık. [0,1] arası değerleri yoğun negatif değerlerine; [1, 2] arası değerleri zayıf negatif değerlerine; [2, 3] arası değerleri nötr değerlere; [3, 4] arası değerleri zayıf pozitif değerlerine; [4, 5] arası değerleri yoğun pozitif değerlerine karşılık düşürüldü. Çalışmamızda kullanılan parametreler yorumların polarite oranları ile ilgilidir; "X_1: Yoğun negatiflik içeren kelimelerin kullanım oranı", "X_2: Zayıf negatiflik içeren kelimelerin kullanım oranı, "X_3: Nötr ifadelerin kullanım oranı", "X_4: Yoğun pozitiflik içeren kelimelerin kullanım oranı" ve "X_5: Zayıf pozitiflik içeren



Şekil 4. Çok Katmanlı Yapay Sinir Ağları modeli (Multi-Layer Perceptron model)

kelimelerin kullanım oranı"dır. Bu durumda doğrusal regresyon modeli aşağıdaki gibi ifade edilebilir.

$$h_i = w_{i1}x_{i1} + w_{i2}x_{i2} + w_{i3}x_{i3} + w_{i4}x_{i4} + w_{i5}x_{i5} + b \quad (7)$$

$$y_k = w_{k1}x_{k1} + w_{k2}x_{k2} + w_{k3}x_{k3} + w_{k4}x_{k4} + w_{k5}x_{k5} + c \quad (8)$$

Burada w değerleri regresyon katsayılarına karşılık gelir. Gizli katmanlardaki nöronların değeri Eş. 7 kullanılarak, çıkış nöronlarının değeri Eş. 8 kullanılarak elde edilir. Veriler daha sonra en küçük kareler kullanılarak doğrusal bir regresyon modelindeki hata hesaplanır.

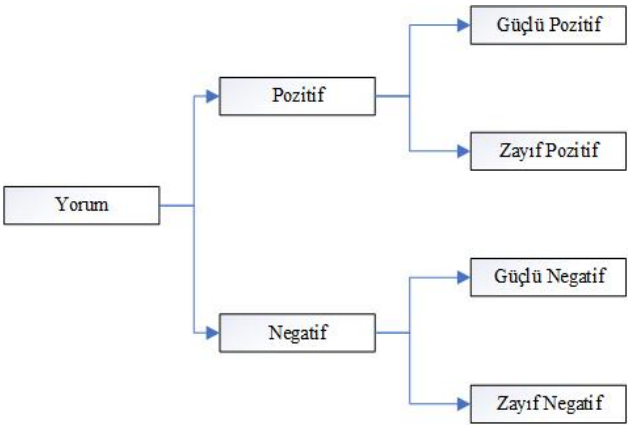
$$\mathrm{E} = \frac{1}{\mathrm{N}}(\sum_{i=1}^{n} t_i - y_i)^2 \quad (9)$$

Doğrusal regresyonda amaç gerçek çıktılar (t_i) ile tahmin edilen çıktılar (y_i) arasındaki toplam farkı yani Eş. 9'da gösterilen hata fonksiyonunu en aza indirgemektir.

3.4. Duygu Analizinde Wavelet Kullanımı
(Document Sentiment Analysis with Wavelet)

Wavelet dönüşümünde anlamlı sonuç elde edebilmek için, hiyerarşik ikili ağaçta her düğümün çocukları birbiriyle kesişmeyen hatta birbirini tamlayan ya da zıt kavramlar olmalıdır. Çünkü Haar dönüşümünden sonra elde edilen vektörün katsayıları bu ikili çocuk düğümlerin hangisinin baskın olduğu hakkında bilgi verir. Eğer ikili çocuk düğümlerin ağırlıkları eşitse ya da ikisi hakkında da bilgi yoksa o ikili sınıfa ait katsayı 0 olur (Bölüm-2'deki hiyerarşik sınıflandırma ağacı örneği gibi).

Önerdiğimiz modelde yorumların terim frekans matrisleri oluşturulup Haar dönüşüm matrisi ile çarpılarak D matrisi oluşturulmakta daha sonra bu D matrisinin ağırlıkları kullanılarak her bir yorumun hangi sınıfta olduğu belirlenmektedir.



Şekil 5. Duygu analizinde kullanılan hiyerarşik sınıf
(Hierarchical classes in sentiment analysis)

Şekil 5'de duygu analizi için kullandığımız hiyerarşik ağaç gösterilmektedir. Modelimizde wavelet ile metin duygu analizi yapılmak için öncelikle olumlu güçlü duygusal, olumlu zayıf duygusal ve olumsuz güçlü duygusal, olumsuz zayıf duygusal yorumların listesine ihtiyaç duyulmaktadır. İngilizce için hazırlanmış bu listeler Pittsburg Universitesinin hazırladığı Öznel Duygu İşaretlemeleri sayfasından alınmıştır [31, 32]. Yorum paragrafları içindeki kelimeler "Güçlü Pozitif Yorum"; "Zayıf Pozitif Yorum"; "Güçlü Negatif Yorum" ve "Zayıf Negatif Yorum" sözlüklerinde aratılıp; bu dört gruptan birine atanmaktadır. Tüm kelimeler aratıldığında, her bir çeşit sınıftan kaç tane sözcük olduğu bulunmaktadır. Her bir yorum için T_f vektörünün birinci elemanı bu yorumdaki güçlü pozitif kelime sayısı; ikinci elemanı zayıf pozitif kelime sayısı; üçüncü elemanı güçlü negatif kelime sayısı; dördüncü elemanı zayıf negatif kelime sayısına eşit olur.

Yorum1: *There's* ***little*** *to* ***recommend*** *Snow Dogs, unless one considers* ***cliched*** *dialogue and* ***perverse*** *escapism a source of high* ***hilarity****.*

Yorum2: *Even fans of Ismail Merchant 's work, I* ***suspect****, would have a* ***hard*** *time sitting through this one.*

Yorum3: *A* ***welcome relief*** *from baseball movies that try too* ***hard*** *to be mythic, this one is a* ***sweet*** *and* ***modest*** *and* ***ultimately*** *winning story.*

Yorum4: *Kung Pow is Oedekerk's realization of his childhood dream to be in a martial-arts flick, and proves that sometimes the* ***dreams*** *of youth should remain just that."*

Yorum5: *Just as Desplechin never allows the audience to understand what it is that fuels Esther to begin her career as an actress, he also keeps her motivations obscured as she continues to develop. When her* ***mentor*** *(Ian Holm) claims she lacks the life experience to express that ever-elusive truth on the stage, she launches herself into what is clearly a* ***self-destructive*** *relationship with a foreign literary* ***critic****, seemingly with the express intention of getting her* ***heart*** *broken.*

Şekil 6. Yorum örnekleri (Example review)

Şekil 6'da eğitim verisinden alınan ve farklı sınıflara ait beş tane yorum paragrafı bulunmaktadır. Bu kısımda bu beş paragraf için T_f matrisinin nasıl oluşturulacağı ve sınıf atamalarının nasıl yapıldığı detaylı bir şekilde anlatılacaktır. Tablo 4'de örnek yorum paragrafları için terim frekans vektörünün içeriğini oluşturan ve yorum polaritesinde etki yaratan terimler gösterilmektedir.

Aynı sınıftan kelimeler bulundukça o sınıfın değeri bir artırılmaktadır. Hiç kelime bulunamadıysa sınıf değeri 0 olarak kalmaktadır. Tablo 4'de sözlüklerde bulunan kelimeler gösterilirken, Tablo 5'de bunların sayısal değerleri yer almaktadır.

Sonuç olarak Tablo 4'daki elemanlar T_f vektörünün elemanlarını oluşturmaktadır. Her bir satır bir yorum paragrafını göstermek üzere yukardaki örnek için T_f, Eş. 10'de gösterilmiştir.

$$T_f = \begin{vmatrix} 2 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \\ 3 & 3 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 2 & 1 & 1 \end{vmatrix} \quad (10)$$

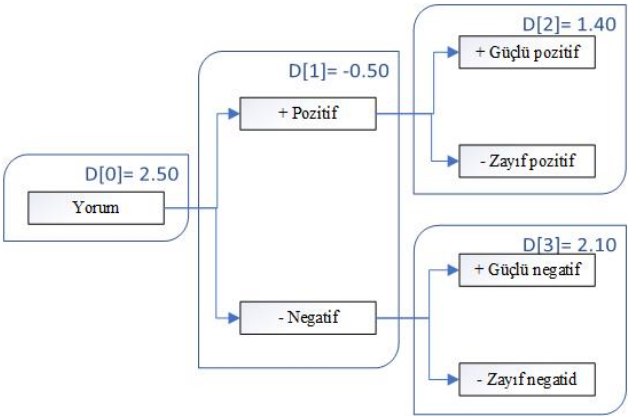
$$H_4 = \begin{vmatrix} \sqrt{\frac{1}{4}} & \sqrt{\frac{1}{4}} & \sqrt{\frac{1}{4}} & \sqrt{\frac{1}{4}} \\ \sqrt{\frac{1}{4}} & \sqrt{\frac{1}{4}} & -\sqrt{\frac{1}{4}} & -\sqrt{\frac{1}{4}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} & 0 & 0 \\ 0 & 0 & \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{vmatrix} \quad (11)$$

Eş. 12'de formülü verildiği üzere; Eş. 10'da değerleri verilen terim frekans matrisi (T_f), Eş. 11'de gösterilen 4x4 boyutlu Haar Dalgacık dönüşüm matrisinin transpozu ile çarpıldığında; baskınlıklarına göre ara sınıfların katsayılarını veren D dönüşüm matrisi elde edilmektedir. D dönüşüm matrisinin eleman değerleri Eş. 13'de gösterilmektedir. Haar dalgacık dönüşüm matrisinin kaça kaçlık bir matris olacağı hiyerarşik ağaçtaki yaprak düğüm sayısına bağlıdır. Haar nxn'deki n sayısı yaprak düğüm sayısına eşittir.

$$D = T_f \; x \; {H_4}^T \quad (12)$$

$$D = \begin{vmatrix} 2{,}5 & -0{,}5 & 1{,}4 & 2{,}1 \\ 1{,}0 & -1 & 0{,}0 & -1{,}4 \\ 3{,}5 & 2{,}5 & 0{,}0 & -0{,}7 \\ 0{,}5 & 0{,}5 & -0{,}7 & 0{,}0 \\ 2{,}0 & 0{,}0 & -1{,}4 & 0{,}0 \end{vmatrix} \quad (13)$$

D dönüşüm matrisinde satırlar paragrafları gösterirken; sütunlar hiyerarşik ağaçtaki sınıfların baskınlık bilgisini verir. Birinci sütun paragraflarda yorum içeren terimlerin frekansı ile orantılıdır. İkinci sütundaki rakamlar o paragrafın genel olarak pozitif mi yoksa negatif mi olduğu bilgisini verir. Eğer rakam 0 ise o paragrafın nötr olduğu bilgisini verir. Üçüncü sutundaki rakamlar pozitif terimler arasında güçlü terimler mi baskın, zayıf terimler mi baskın yoksa ikisi eşit ağırlıkta mı bilgisini verir. Son olarak dördüncü sutun negatif terimlerden güçlü terimler mi zayıf terimler mi baskın bilgisini verir.



Şekil 7. D matrisindeki ağırlıklar ve üst sınıfların ilişkisi (The weights of D matrix and upper classes relation)

Birinci yorumda 2 güçlü pozitif, 3 güçlü negatif terim vardır. Güçlü terimler zayıf terimlere baskındır. Negatif terimler ise pozitif terimlere baskındır. Şekil 7'de D vektörüne baktığımızda güçlü zayıf baskınlıklarını veren D[2] ve D[3] pozitif değer alırken; D[1] negatif değer almaktadır.

Tablo 4. Yorum paragrafından T_f vektörünün oluşturulma aşaması (Extracting T_f vector from review)

	Yorum1	Yorum2	Yorum3	Yorum4	Yorum5
Pozitif güçlü terimler	Recommend, hilarity	-	Welcome, relief, ultimately	-	-
Pozitif zayıf terimler	-	-	Welcome, sweet, modest	dream	Mentor, heart
Negatif güçlü terimler	Little,cliched, perverse	-	-	-	Self-destructive
Negatif zayıf terimler	-	suspect, hard	hard	-	Critic

Tablo 5. T_f eleman değerleri (T_f values)

	T_f -1	T_f -2	T_f -3	T_f -2	T_f -3
Pozitif güçlü terimler	2	0	3	0	0
Pozitif zayıf terimler	0	0	3	1	2
Negatif güçlü terimler	3	0	0	0	1
Negatif zayıf terimler	0	2	1	0	1

Duygu sınıflandırma probleminde amacımız paragraflarda bulunan yorumların duygu sınıflarından ($C_{güçlü\_olumsuz}$, $C_{zayıf\_olumsuz}$, $C_{nötr}$, $C_{güçlü\_olumlu}$, $C_{zayıf\_olumsuz}$) hangisine ait olduğunu bulmaktır. Öncelikle yorum 1'e ait T_f vektörü wavelet dönüşümü ile D vektörü (D) haline gelir. Wavelet yönteminde vektörün hangi sınıfa daha yakın olduğu bulunurken D vektörüne ait katsayılar kullanılır. D[i][j] dönüşüm matrisinde i'ler satırları yani paragrafları; j'ler polarite sınıflarını gösterir. D[0][j]'de i=0 olması, bu vektörün birinci paragraf yorumuna ait olduğunu gösterir.

D'nin ikinci sütunu (1 indeksli sütun) sıfırdan büyük ise yorumun genel olarak pozitif; küçük ise genel olarak negatif olduğu söylenir. D'nin ikinci sütunu 0 ise pozitif terimlerin negatif terimlere eşit olduğu yada bu terimlerin hiç olmadığı anlamına gelir; D'nin üçüncü sütunu (2 indeksli sütun) işaretine göre; yorumun genel olarak güçlü pozitif, zayıf pozitif yada nötr olduğu bilgisini verir; D'nin dördüncü satırı (3 indeksli satır) işaretine göre yorumun genel olarak güçlü negatif, zayıf negatif yada nötr olduğu bilgisini verir.

Çalışmamızda, D dönüşüm vektörünün elemanları her bir sınıfın baskınlık değeriyle ilgilidir. Her bir paragraf yani satır için; eğer bir indeksli satır nötr ise yorumun nötr sınıfında olduğu söylenebilir. Eğer bir indeksli sütun olumlu (pozitif) ise o zaman güçlü olumlu ya da güçsüz olumlu olduğunu anlamak için iki indeksli sütuna bakılır. Burada, iki indeksli sütundaki eleman negatif ise $C_{zayıf\_olumlu}$, pozitif ise $C_{güçlü\_olumlu}$ sınıfına daha yakındır. Tam tersine, eğer bir indeksli sütun olumsuz (negatif) ise o zaman güçlü olumsuz ya da güçsüz olumsuz olduğunu anlamak için üç indeksli sütuna bakılır. Burada yorum paragrafı, üç indeksli satırdaki eleman negatif ise $C_{zayıf\_olumsuz}$, pozitif ise $C_{güçlü\_olumsuz}$ sınıfına atanır. D'nin elemanlarının mutlak değeri büyüdükçe bir sınıfın diğer sınıfa baskınlığının arttığını söyleyebiliriz. Şekil 8'den görüldüğü üzere D eleman değerlerine göre yorumların sınıflarına karar verilir. Eş. 14 ve Eş. 15'de görülen katsayılar, wavelet dönüşümü sonucunda çıkan yorum vektör değerleri kullanılarak elde edilmiştir. Çalışmamızda W katsayıları sadece baskınlığın belirgin olan yorumlarda; wavelet'in kullanılmasını sağlamak için bir eşik değeri olarak kullanılmaktadır.

$$W_{pos} = 2^{|D[i][1]|} * 2^{|D[i][2]|} \quad (14)$$

$$W_{neg} = 2^{|D[i][1]|} * 2^{|D[i][3]|} \quad (15)$$

Burada 2 tabanlı üssel bir matematik ifadesi kullanılmasının nedeni D[i][1] mutlak değeri 0'dan büyük herhangi bir değer aldığında sınıfa girme ihtimalinin olmasıdır yani D[i][1] sayısı 0,005 gibi küçük bir sayı olsa bile üssel bir biçimde yazıldığında w değeri 0 dan büyük (1,005) olacaktır. Ayrıca D[i][1] mutlak değeri ne kadar büyürse w değeri de büyüyeceğinden o sınıfın seçilme ihtimali artacaktır.

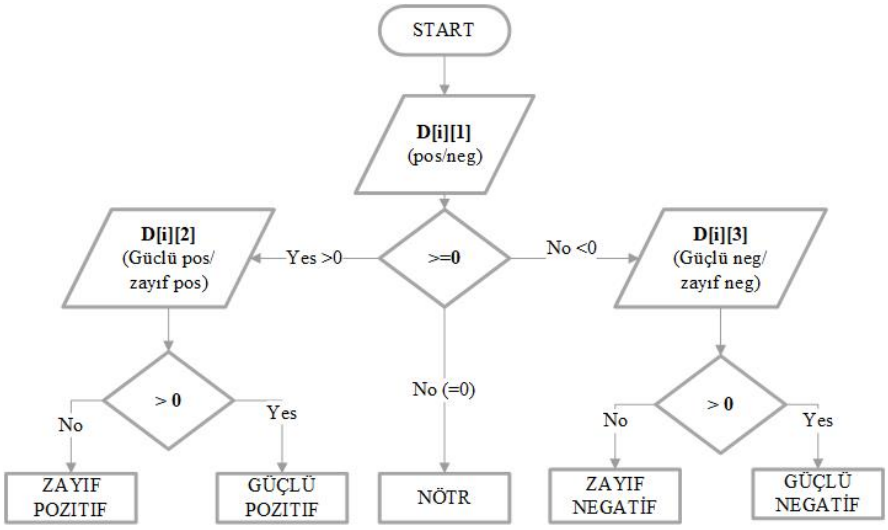
Şekil 6'daki örnek yorumlar için T_f matrisi Haar dönüşüm matrisinin transpozu ile çarpılarak Eş. 13'deki D matrisi elde edilmişti. D matrisi ile Eş. 14 ve Eş. 15 kullanılarak Eş. 16'da gösterilen çıkış matrisi O bulunmuştur. Bu matris yorumlara önerilen sınıfları ve sınıf ağırlıklarını içermektedir.

$$O = \begin{bmatrix} -2 & 13{,}75 \\ -1 & 11{,}18 \\ 2 & 12{,}18 \\ 1 & 3{,}34 \\ 0 & 2 \end{bmatrix} \quad (16)$$

D dönüşüm matrisinin bulunmasından başlayıp sınıfların ve sınıf ağırlıklarının atanmasına kadar geçen süreç ile ilgili kod parçası Şekil 9'de gösterilmektedir.

Wavelet ile Harmanlanmış Sınıflandırma Algoritması Adımları

- Adım.1: Kısa metinlerden oluşan yorumlar birer paragraf gibi değerlendirilip öncelikle her bir yorum paragrafı için terimler "güçlü pozitif", "zayıf pozitif", "güçlü negatif", "zayıf negatif" sözlüklerinde aratılıp 4 boyutlu terim frekans vektörü bulunmuştur.



Şekil 8. Dönüşüm vektörüne göre sınıfların karar verilmesi akış diyagramı (Classes according to transformation vector)

- Adım.2: 4 sınıftan oluşan bu terim frekans vektörü 4x4'lük Haar Dönüşüm matrisinin devriği ile çarpılarak D dönüşüm matrisi elde edilmiştir. Bu matrisin satırları paragrafları sütunları ise sınıflardan hangisinin baskın olduğu ile ilgilidir.
- Adım 3: D dönüşüm matrisi okunup, değerlendirilerek her bir yorum paragrafının hangi sınıfa ait olduğu Şekil 11'deki akış diyagramındaki gibi bulunur ve baskınlığın ölçüsünü belirleyen w değeri hesaplanır.
- Adım 4: Daha sonra w'nin belli bir eşik değerinin üstünde olduğu baskınlığın güçlü olduğunu bildiğimiz paragraflar wavelet'in bulduğu sınıfa atanır. Burada wavelet yöntemi son bulmaktadır.
- Adım 5: Geri kalan yorum paragraflarında içinde bulunan her kelime dikkate alınarak terim frekans vektörleri oluşturulup (Buradaki T_f'lerde gereksiz sözcükler dışındaki tüm sözcükler dikkate alınmaktadır) bilinen sınıflandırma algoritmaları paragraflara uygulanmıştır.

4. DEĞERLENDİRME (EVALUATION)

Bu bölümde, önerdiğimiz yöntem değerlendirilmekte; yorum metinlerinin görselleştirilmesi ve wavelet yönteminin klasik algoritmalarla birleştirildiğinde doğrulukların artışı ile ilgili sonuçlar gösterilmektedir.

4.1. Metin Duygusunun Görselleştirilmesi
(Document Sentiment Visualation)

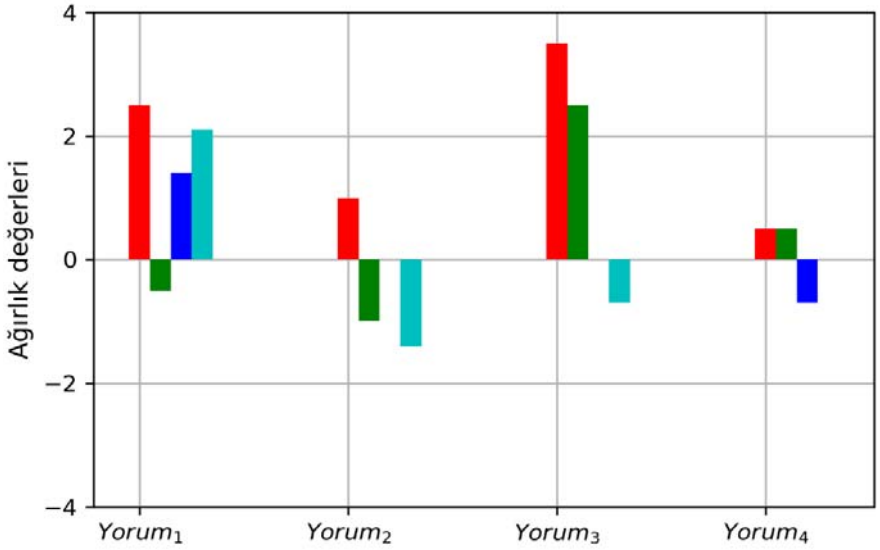
Her bir yoruma ait terim frekans vektörüne wavelet dönüşümü uygulanılarak dönüştürülür. Yeni elde edilen matris her bir yorum için 4 farklı polarite sınıfına ait katsayılar içermektedir. Şekil 10, D matrisinin ilk 4 satırı için çizge grafik ile gösterilmiş halidir. Kırmızı sütunlar o paragrafta ne kadar yorum içeren sözcük olduğu hakkında bilgi içerir.

Her bir yorum için ilk sütun paragrafın genel olarak pozitif ya da negatif içerikli olduğu yada eğer iki yorumdan da eşit sayıda ise nötr olduğu hakkında bilgi verir. İkinci sütun 0'ın altında değer alıyorsa yorumlar genel olarak negatif, üstünde değer alıyorsa yorum genel olarak pozitif demektir. Yorum 1 ve yorum 2 genel olarak negatif içerikli olmasına karşın, yorum 3 ve yorum 4 pozitif içeriklidir. Üçüncü sütun pozitif yorumlarda güçlü yorumların mı zayıf yorumların mı baskın olduğu bilgisini verir. Görüldüğü üzere yorum 1'de güçlü pozitif yorumlar baskınken, yorum 4'de zayıf pozitif yorumlar baskındır. Diğer iki yorumda bu değer 0'dır. Her bir yorum için dördüncü sütun negatif yorumlarda güçlü yorumların mı zayıf yorumların mı baskın olduğu bilgisini verir.

```
# -2:strong negative; # -1:weak negative; # 0:nötr
# 1:weak positive; # 2:strong positive
import math as m
#outputda her satır için ilk eleman "sınıfı" ikinci eleman "ağırlığı" vermektedir
output = [[99] * 2 for i in range(datalength)]
thresholdNum=0
for i in range(datalength):
    if(D[i][1]==0):
        w=2
        if w>thresholdNum:
            output[i][1]="2"#Nötr ağırlığı
            output[i][0]="0"#Nötr sınıfında
    if(D[i][1]>0):
        if(D[i][2]>=0):# Strong pozitif sınıfında
            w=pow(m.e,D[i][1])*pow(m.e, D[i][2]) #ağırlık
            if w>thresholdNum:
                output[i][1]=w
                output[i][0]="2"
        else: # Week pozitif sınıfında
            w=pow(m.e,D[i][1])*pow(m.e, (-1)*D[i][2]) #ağırlık
            if w>thresholdNum:
                output[i][1]=w
                output[i][0]="1"
    if(D[i][1]<0):
        if(D[i][3]>=0): # Strong negatif sınıfında
            w=pow(m.e,(-1)*D[i][1])*pow(m.e, D[i][3]) #ağırlık
            if w>thresholdNum:
                output[i][1]=w
                output[i][0]="-2"
        else:# Week negatif sınıfında
            w=pow(m.e,(-1)*D[i][1])*pow(m.e, (-1)*D[i][3]) #ağırlık
            if w>thresholdNum:
                output[i][1]=w
                output[i][0]="-1"
print (output)
```

```
[['-2', 13.75], ['-1', 11.18], ['2', 12.18], ['1', 3.34], ['0', '2']]
```

Şekil 9. Örnek 5 yorum satırı için D Haar Dönüşüm Matrisi ile çıkış sınıflarının ve ağırlıkların hesaplanması
(Finding the output classes and relative weights for 5 review examples)



Şekil 10. Metin duygu analizi çizge grafik gösterimi (Grraphical representation of text sentiment classification)

4.2. Wavelet'in Duygu Sınıflandırma Algoritmalarına Katkısı (The wavelet contribution for sentiment classification)

Bu bölümde değişik makina öğrenmesi yöntemleri, wavelet duygu analiz yöntemi ile birleştirilerek doğruluklar test edilmiş sonuçlar klasik yöntemlerle kıyaslanmıştır.

Wavelet dönüşüm fonksiyonunun duygu analizindeki etkisini test etmek için UCI veri tabanındaki iki farklı veri seti kullanılmıştır. İlaç yorumları ve IMDB film yorumlarını içeren bu iki veri setinde [33-36] ilaç yorumları için 215063; IMDB için 3000 tane yorum verisi bulunmaktadır. İlaç yorumlarında yorumun pozitif ya da negatif olduğu 1'den ona kadar puanlandırılmıştır. Beş puanı nötr'e karşılık gelmektedir. Bir çalışmamızda puanları gruplayıp 1,2 puanlarına güçlü negatif; 3,4 puanlarına negatif; 5 nötr; 6-7,8 pozitif ve 9,10 güçlü pozitif olarak değiştirdik. IMDB veri setinde sonuç sınıflar işaretli olup; nötr, güçlü pozitif, zayıf pozitif, güçlü negatif ve zayıf negatif şeklindeydi.

Çalışmamız jupyter notebook üzerinde math, numpy, pandas, scikit-learn gibi python kütüphaneleri kullanılarak kodlanmıştır. Çalışmamızın özgün kısmı olan, Haar Dönüşümü ile sınıflandırma yapan kod parçası, eğitim verisinin küçük bir kısmı için çalıştırılıp eklenmiştir (Şekil 9). Klasik algoritmalar ve wavelet yöntemi ile yapılan duygu sınıflandırmasındaki doğrulukları hesaplamak için Tablo 6 ve Tablo 7'deki ölçekler kullanılmıştır.

Her algoritmada her sınıf için doğruluk değerleri ayrı ayrı bulunmuş, daha sonra doğrulukların ortalama değerleri alınmıştır. Burada wavelet çözümünde bir gözlemin sonucunu wavelet'in belirlemesini kabul etmek için minimum eşik değeri (eşik değeri) 2 ve 4 olarak kabul edilmiştir.

Tablo 8 ve Tablo 9'de İlaç yorumlarının sınıflandırma sonuçları yer almaktadır. Tablo 8'de ilaç yorumları için klasik algoritmalar ve wavelet dönüşüm uygulamamız ayrı ayrı değerlendirilmektedir. Bu tablodan görüleceği üzere wavelet tek başına kullanıldığında klasik algoritmalardan daha iyi sonuç vermemektedir. Tablo 9'da wavelet katsayıları {2, 4} eşik değerinden büyükse yani yeterince güçlü bir şekilde polarite içeriyorsa; wavelet sonuçları o verinin hangi sınıfta olacağını belirlemiştir. Eşik değerinin altında kalan örnekler için klasik algoritmalar uygulanmıştır. Yani wavelet yöntemimiz klasik yöntemlerle harmanlanmıştır. Klasik algoritmalarla birleştirilmiş, eşik değeri 4'den büyükken ve eşik değeri 2'den büyükken elde edilen doğruluklar için Tablo 9'a bakılabilir. Görüldüğü üzere klasik yöntemlerle birlikte parçacık dönüşüm metodunun kullanılması doğrulukları %3-5 oranında artırmıştır.

Tablo 6. Kullanılan Ölçüm Metrikleri-I (Evaluation Metrics-I)

	Gerçek pozitif gözlemler	Gerçek negatif gözlemler
Pozitif olarak tahmin edilenler	TP (Doğru olan pozitif tahminlerin sayısı)	FP (Yanlış olan pozitif tahminlerin sayısı)
Negatif olarak tahmin edilenler	FN (Yanlış olan negatif tahminlerin sayısı)	TN (Doğru olan negatif tahminlerin sayısı)

Tablo 7. Kullanılan Ölçüm Metrikleri -II (Evaluation Metrics-II)

Doğruluk ölçütü	Formülü
Doğruluk (Accuracy)	(TP+TN)/(TP+TN+FP+FN)
Hassasiyet (Precision)	TP/(TP+FP)
GeriÇağırma (Recall)	TN/(TN+FN)
F-ölçütü	2* Hassasiyet* GeriÇağırma/(Hassasiyet+GeriÇağırma)

Tablo 8. İlaç yorumları veri seti ile duygu analizi sonuçları
(The results for medical reviews)

Algoritma	P	R	F	Acc
Linear Regression	0,87	0,69	0,76	0,88
SVM	**0,93**	**0,84**	**0,88**	**0,92**
Multi Layer Perceptron	0,96	0,58	0,72	0,86
Wavelet (w≥2)	0,54	0,59	0,56	0,56
Wavelet (w≥4)	0,68	0,72	0,69	0,68

Tablo 9. İlaç yorumları için wavelet ile birleştirilmiş duygu analizi sonuçları
(The results of hybrid wavelet method for medical reviews)

Algoritma	P	R	F	Acc
Wavelet (w≥2)+Linear Reg,	0,80	0,82	0,80	0,88
Wavelet (w≥4)+Linear Reg,	0,83	0,92	0,87	0,91
Wavelet (w≥2)+SVM	0,91	0,91	0,91	0,93
Wavelet (w≥4)+SVM	**0,94**	**0,97**	**0,95**	**0,97**
Wavelet (w≥2)+ MLP	0,92	0,63	0,74	0,86
Wavelet (w≥4)+ MLP	0,93	0,70	0,79	0,89

Tablo 10. IMDB film yorumları veri seti ile duygu analizi sonuçları
(The results for movie reviews)

Algoritma	P	R	F	Acc
Linear Regression	0,80	0,72	0,75	0,75
SVM	**0,89**	**0,82**	**0,85**	**0,85**
Multi Layer Perceptron	0,85	0,79	0,81	0,81
Wavelet (w≥2)	0,63	0,69	0,65	0,65
Wavelet (w≥4)	0,73	0,81	0,76	0,77

Tablo 11. IMDB film yorumları için wavelet ile birleştirilmiş duygu analizi sonuçları
(The results of hybrid wavelet method for movie reviews)

Algoritma	P	R	F	Acc
Wavelet (E≥2)+Linear Reg,	0,84	0,76	0,79	0,80
Wavelet (E≥4)+Linear Reg,	0,89	0,84	0,86	0,87
Wavelet (E≥2)+SVM	0,90	0,84	0,86	0,86
Wavelet (E≥4)+SVM	**0,91**	**0,86**	**0,88**	**0,88**
Wavelet (E≥2)+ MLP	0,87	0,81	0,83	0,83
Wavelet (E≥4)+ MLP	0,89	0,84	0,86	0,86

Tablo 10 ve Tablo 11'de IMDB film yorumları için tek tek ve hibrit doğrulukları göstermektedir. İlaç yorumlarına benzer şekilde wavelet'in klasik yöntemlerle birleştirilmesi sonuçları %3 artırmıştır.

5. SONUÇLAR VE TARTIŞMALAR (RESULTS AND DISCUSSIONS)

Wavelet her ne kadar sinyal işleme alanında ortaya çıkan bir metot olsa da, çalışmamız sonucunda hiyerarşik sınıfların ağırlıklandırılmasında, doküman sınıflandırmada ve bir çeşit sınıflandırma olan duygu analizinde fayda sağladığı görülmektedir. Çalışmamız duygu analizi uygulamasında wavelet'i eski yöntemlerle birlikte kullanarak doğruluğun %3-5 oranında artığını göstermektedir. 2020'de yayınlanan yeni bir çalışmada CNN-LSTM yöntemi kullanılarak UCI ilaç yorumları veri seti sınıflandırılmaktadır. Bu çalışmada son derece güncel algoritmalar kullanılmakla birlikte 92,24'lük bir başarı elde edilmiştir [34]. Ki bu başarı halen SVM algoritmasının Wavelet ile birleştirilerek kullanıldığı haldeki 0,97'lik başarıdan küçüktür. Çalışmanın öne çıkan sonuçları aşağıdaki gibi sıralanabilir.

- Bir metin içindeki her bir sütunun duygu polaritesi wavelet dönüşümü kullanılarak görselleştirilebilir. Bu amaçla daire grafiği ve çubuk grafiği kullanılabilir.
- Wavelet tek başına kullanıldığında yeterince yüksek doğruluklar vermese de klasik algoritmalarla birleştirildiğinde doğrulukların artmasına neden olmuştur.
- İlaç yorumları veri seti üzerinde yapılan duygu sınıflandırması probleminde; SVM ile 0,92 doğruluk elde edilirken, SVM'e wavelet de eklendiğinde doğruluk 0,97'ya çıkmıştır.
- Film yorumları veri seti üzerinde yapılan duygu sınıflandırması probleminde; SVM ile 0,85 doğruluk elde edilirken, SVM'e wavelet de eklendiğinde doğruluk 0,88'e çıkmıştır.

Bu çalışmanın metnin içeriklerinin hiyerarşik sınıflandırıldığı web uygulamalarında ya da numerik dahi olsa metinlerin ikili ağaç sınıflarından oluştuğu veri örneklerinin sınıflandırma problemlerinde fayda sağlayacağı umulmaktadır. Biz, bundan sonraki çalışmamızda veri setinin, yaprak düğümlerin sayısal ve ikili ayrık sınıflara ait yapıda olduğu alanlarda wavelet dönüşümü kullanmayı ve sonuçları gözlemlemeyi düşünüyoruz.

6. SONUÇ LAR (CONCLUSIONS)

Çalışmamız wavelet'i şimdiye kadar çok az sayıda çalışma yapılmış bir alanda; dokümanların analiz ve görselleştirmesinde kullanmaktadır. Wavelet kelimelerin hiyerarşik üst anlam ağaçları içindeki üst sınıflara (üst terimlere) ait ağırlıkların bulunmasında, dokümanların hangi sınıfa ait olduğunun çıkarılmasında, doküman içeriklerinin görselleştirilmesinde ve duygu analizi uygulama alanlarında fayda sağlamaktadır. Çalışmamızda wavelet klasik yöntemlere destek verecek şekilde kullanıldığında, doğrulukların arttığı görülmüştür.

TEŞEKKÜRLER (ACKNOWLEDGEMENT)

Yazarlar, wavelet uygulamaları konusundaki araştırmalarımıza sağladığı önemli katkılar ve desteklerden dolayı ISIAM (Hindistan Endüstri ve Uygulamalı Matematik Topluluğu) Başkanı olan ve maalesef 20 Ocak 2020 tarihinde aramızdan ayrılan Prof. Dr. Abul Hassan SIDDIQI'yı saygıyla anarlar ve çalışmalarını Prof. Dr. Abul Hassan SIDDIQI'ya ithaf ederler.

KAYNAKLAR (REFERENCES)

1. Akansu A.N. ve Smith M.J., Subband and wavelet transforms: design and applications, Springer Science & Business Media, 340, 2012.
2. Meyer Y. ve Ryan, R.D., Wavelets: Algorithms and Applications. SIAM, Philadelphia, PA, 1993.

3. Chan Y.T., Wavelet basics. Springer Science & Business Media, 1994.

4. Strang G. Ve Nguyen, T., Wavelets and Filter Banks,Wellesley-Cambridge Press, ISBN 0-9614088-7-1 Box 812060, Wellesley MA 02181 USA, 1996.

5. Daubechies, I., Ten Lectures on Wavelets, Society for Industrial and Applied Mathematics, Capital City Press, Philadelphia, Pennsylvania, 1992.

6. Daubechies I., The wavelet transform, timefrequency localization and signal analysis. IEEE Trans. Inf. Theor., 36 (5), 961–1005, 2006.

7. Farkov Y. A., P. Manchanda A. H. ve Siddiqi, Construction of Wavelets Through Walsh Functions, ISBN - 978 -981-13-6370-2 (e-book), 382, 2019.

8. Polikar R., The story of wavelets. Physics and modern topics in mechanical and electrical engineering, 192-197, 1999.

9. Siddiqi A. H., Manchanda, P. ve Kocvara, Fast wavelet-based algorithms for option pricing. In Proc. world Multi conference on Systemic, Cybernetics and Informatics, 2002.

10. Li T., Li Q., Zhu S. ve Ogihara, M., A survey on wavelet applications in data mining. ACM SIGKDD Explorations Newsletter, 4 (2), 49-68, 2002.

11. Aggarwal C. C., On effective classification of strings with wavelets. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 163-172, 2002.

12. Xu C. ve Zhou Y. M., Wavelet-based hierarchical document categorization. In 2007 International Conference on Wavelet Analysis and Pattern Recognition IEEE, 4, 1524-1527, 2007.

13. Xexéo G., de Souza J., Castro P. F. ve Pinheiro W. A., Using wavelets to classify documents. In 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology IEEE, 1, 272-278, 2008.

14. Mahajan A., Jat S., ve Roy S., Feature Selection for Short Text Classification using Wavelet Packet Transform. In Proceedings of the Nineteenth Conference on Computational Natural Language Learning, 321-326, 2015.

15. Alasmari S. F. ve Dahab, M., Sentiment detection, recognition and aspect identification. International Journal of Computer Applications, 975, 8887, 2017.

16. Bollen J., Mao H., ve Zeng X., Twitter mood predicts the stock market. Journal of computational science, 2 (1), 1-8, 2011.

17. Taboada M., Brooke J., Tofiloski M., Voll, K. ve Stede, M., Lexicon-based methods for sentiment analysis. Computational linguistics, 37 (2), 267-307, 2011.

18. Cambria E., Schuller B., Xia, Y. ve Havasi C., New avenues in opinion mining and sentiment analysis. IEEE Intelligent systems, 28 (2), 15-21, 2013.

19. Thelwall M., Buckley K., ve Paltoglou G., Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology, 63 (1), 163-173, 2012.

20. Kilimci Z.H., Financial Sentiment Analysis with Deep Community Models for Stock Market (DCM), Journal of the Faculty of Engineering and Architecture of Gazi University, 35 (2), 635-650, 2020.

21. Çoban Ö., Özyer B., ve Özyer G. T., Sentiment analysis for Turkish Twitter feeds. In 2015 23nd Signal Processing and Communications Applications Conference (SIU) IEEE, 2388-2391, 2015.

22. Chidambarathanu K. ve Shunmuganathan K. L., Predicting user preferences on changing trends and innovations using SVM based sentiment analysis. Cluster Computing, 1-5, 2017.

23. Zhang W., Kong S. X. ve Zhu Y. C., Sentiment classification and computing for online reviews by a hybrid SVM and LSA based approach. Cluster Computing, 22 (5), 12619-12632, 2019.

24. Akhtar M. S., Kumar A., Ghosal D., Ekbal A. ve Bhattacharyya P., A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 540-546, 2017.

25. Alboaneen D. A., Tianfield H. ve Zhang Y., Sentiment analysis via multi-layer perceptron trained by meta-heuristic optimisation. In 2017 IEEE International Conference on Big Data, 4630-4635, 2017.

26. Jotheeswaran J. ve Koteeswaran S., Decision tree based feature selection and multilayer perceptron for sentiment analysis. Journal of Engineering and Applied Sciences, 10 (14), 5883-5894, 2015.

27. Dang N. C., Moreno-García M. N. ve De la Prieta F., Sentiment analysis based on deep learning: A comparative study. Electronics, 9 (3), 483, 2020.

28. Taddy M., Multinomial inverse regression for text analysis. Journal of the American Statistical Association, 108 (503), 755-770, 2013.

29. You Q., Luo J., Jin H. ve Yang J., Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In Proceedings of the Ninth ACM international conference on Web search and data mining 13-22, 2016.

30. Zhang Z., Zou Y. ve Gan, C., Textual sentiment analysis via three different attention convolutional neural networks and cross-modality consistent regression. Neurocomputing, 275, 1407-1415, 2018.

31. Akkaya C., Wiebe J. ve Mihalcea R., Subjectivity word sense disambiguation. In Proceedings of the 2009 conference on empirical methods in natural language processing 190-199, 2009.

32. Multi perspective question answering, Subjectivity sense annotations, http://mpqa.cs.pitt.edu/lexicons/subj_sense_ annotations/,Yayın tarihi: Ağustos 7, 2009, Erişim tarihi: Aralık 16, 2020.

33. Felix G, Surya K, Hagen M ve Sebastian Z. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In Proceedings

of the 2018 International Conference on Digital Health, ACM, New York, NY, USA, 121-125, 2018.

34. UCI Machine Learning Repository, İlaç yorum Dataseti, https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29, Yayın tarihi: Ekim 4, 2018, Erişim tarihi: Aralık 16, 2020.

35. Nagesh H., Goil S. ve Choudhary A.N., UCI Machine Learning Repository, Adaptive Grids for Clustering Massive Data Sets. Department of Energy ASCI, 2020.

36. UCI Machine Learning Repository, Filim yorum Dataseti https://archive.ics.uci.edu/ml/datasets/Movie, Yayın tarihi: Temmuz 7, 1999, Erişim tarihi: Aralık 16, 2020.