DOI: 10.25092/baunfbed.654022 *J. BAUN Inst. Sci. Technol., 21(2), 822-839, (2019)*

Danışmanlı ve yarı danışmanlı öğrenme kullanarak doküman vektörleri tabanlı tweetlerin duygu analizi

Metin BİLGİN[1,*], İzzet Fatih ŞENTÜRK[2]

[1]Bursa Uludağ Üniversitesi Müh. Fak. Bilgisayar Müh. Böl.,Görükle kampüsü, Bursa.
[2]Bursa Teknik Üniversitesi Müh. ve Doğa Bil.Fak. Bilgisayar Müh. Böl., Mimar Sinan Kampüsü, Bursa.

Geliş Tarihi (Received Date): 06.03.2019
Kabul Tarihi (Accepted Date): 10.07.2019

Özet

İnternetin günlük hayatımızdaki artan kullanımı ile beraber sosyal medya organlarının gelişimi de paralellik göstermektedir. Mikroblog adı verilen facebook ve twitter benzeri uygulamaları ile anlık duyguları ve düşünceleri ifade etmek son derece yaygın bir hale gelmiştir. Mikroblog sitelerinin en yaygın kullanıma sahip olanlarından birisi de Twitter uygulamasıdır. Twitter üzerinden paylaşılan mesajlar bir ürün ya da hizmet hakkında olabileceği gibi bir kişiyle ilgili bir yorumda olabilmektedir. Yapılan yorumun belirtmek istediği anlamı ve duyguyu belirleyebilmek son dönemdeki gözde konulardan biridir. Bir ürün ya da hizmet hakkında yapılan binlerce yorumun tek tek okunup anlamlandırılması ve yorumlayanların fikirlerinin sınıflandırılması geleneksel yöntemlerde oldukça zaman ve emek alan bir alandır. Gerek makine öğrenmesi ve derin öğrenme algoritmalarındaki gelişmeler gerekse de bunları işleyip yorumlayacak bilgisayar sistemlerinin gelişimine parallel olarak milyonlarca veri üzerinde duygu sınıflandırılması mümkün hale gelmiştir. Gerçekleştirdiğimiz çalışmada Türkçe ve İngilizce tivitler üzerinde duygusal sınıflandırma çalışması gerçekleştirilmiştir. Döküman vektörleri (Doc2Vec) kullanılarak yapılan çalışmada hem DBoW ve DM gibi iki farklı döküman vektörü yönteminin çalışması hemde Yarı Danışmanlı ve Danışmanlı öğrenmenin etkileri araştırılmıştır. Çalışma sonuçları doğruluk, kesinlik, anma, özgünlük ve F-ölçütü metrikleri ile raporlanmıştır. Gerçekleştirilen çalışma sonucunda Yarı Danışmanlı öğrenme yöntemi hem Türkçe hemde İngilizce veri kümesinde Danışmanlı öğrenmeye göre daha başarılı sonuçlar elde etmiştir.

Anahtar kelimeler: *Yarı danışmanlı öğrenme, danışmanlı öğrenme, Doc2Vec, duygu analizi, makine öğrenmesi, doğal dil işleme, derin öğrenme.*

[*] Metin BİLGİN, metinbilgin@uludag.edu.tr, https://orcid.org/0000-0002-4216-0542
İzzet Fatih ŞENTÜRK, izzet.senturk@btu.edu.tr, https://orcid.org/0000-0002-1550-563X

Sentiment analysis of tweets based on document vectors using supervised learning and semi-supervised learning

Abstract

The increasing presence of the Internet in the daily life leads to proliferation of social media. Microblogging applications such as Facebook and Twitter let their users to express their feelings and emotions in a short form of digital content including text, picture, and video. Twitter is one of the most popular microblogging applications. Tweets posted on twitter may contain positive or negative comments on a product, service or an individual. The analysis of the sentiments behind the tweets has been a popular research topic recently. Considering the time required to read and analyze a vast amount of tweets regarding a given product or service, manual classification of tweets is a complex and time consuming process. Recent machine learning and deep learning algorithms and the developments in the computer systems enable classification of the huge amount of data according to their sentiments in parallel. In this study, we have carried out sentiment analysis on tweets with both Turkish and English content. We have employed Document Vectors (Doc2Vec) along with DBoW and DM and analyzed the performance of semi-supervised and supervised learning methods. We have reported the performance in terms of accuracy, precision, recall, specificity and F-score metrics. We have concluded that semi-supervised learning outperforms supervised learning for both Turkish and English datasets.

Keywords: *Semi-supervised learning, supervised learning, Doc2Vec, sentiment analysis, machine learning, natural language processing, deep learning.*

1. Giriş

Son yıllarda internete erişim imkanlarının artışı ile internet hayatın vazgeçilmez bir nesnesi haline gelmiştir. Özellikle mobil tabanlı sosyal medya uygulamalarının cep telefonlarına girmesiyle hayatımızın her anı sanal ortama taşınmaktadır. İnsanların herhangi bir ürün ya da hizmet ile ilgili görüşlerini hızlı ve kolay bir şekilde paylaşmalarına imkan sağlayan sosyal platform (Facebook, Twitter vb.) siteleri günümüzde çok revaçtadır. İnternetin kullanımının yaygınlaşmasıyla beraber sosyal platformalarda (Twitter, Facebook vb.) paylaşılan içeriklerin artışı paralellik göstermektedir [1]. Sosyal platformlar üzerinden yazılan metinler psikoloji, veri madenciliği gibi araştırma alanları için büyük bir veri toplama alanı oluşmasını sağlamıştır [2,3]. İnternet kullanımının artışı ile birlikte artan kullanıcı yorumlarının olumlu ya da olumsuz gibi sınıflandırılması büyük bir ihtiyaç haline gelmiştir.

Büyük miktarda verinin ön işlemlerden geçirilip, işlenmesi ve doğru sınıflandırılması büyük önem arz etmektedir. Sosyal paylaşım siteleri üzerinden elde edilen veriler üzerinde çeşitli makine öğrenmesi algoritmaları çalıştırılarak sınıflandırmayı gerçekleştirmek mümkündür. Makine Öğrenmesi algoritmalarının yüksek doğrulukla sınıflandırma yapabilmesi için ön işlem aşamasının doğru bir şekilde organize edilmesi gerekmektedir. Sosyal ağlardan elde edilen veriler ise yanlış yazılmış kelimeleri, kısaltmalar ve günlük konuşma diliyle yazılmış kelimeleri ve cümleleri barındırdığından

üzerinde duygu analizi yapmak oldukça zordur. Bu zorlukları aşabilmek için doğal dil işlemenin metodlarından faydalanmak kaçınılmaz bir durum oluşturmaktadır [4].

Bu kısımda önceki çalışmalar hakkında bilgiler verilecektir. Makalenin bu bölümü şu şekilde organize edilmiştir; duygu analizi üzerine yapılmış çalışmalar, Twitter verileri üzerindeki çalışmalar, Danışmanlı Öğrenme kullanan çalışmalar, Yarı Danışmanlı Öğrenme kullanan çalışmalar ve Türkçe Twitter verileri üzerine yapılmış çalışmaları açıklamak şeklindedir.

Günümüzce Twitter verileri üzerinde doğal dil işleme ile ilgili birçok çalışma gerçekleştirilmiştir. Bu çalışmalara örnek olarak: Szomsor ve diğ. [5], belirli bir tarih aralığında paylaşılan tivit metinlerinden salgın hastalıkları tahmin etmeye çalışan bir sistem geliştirmişlerdir. Bian ve diğ. [6] yaptıkları çalışma ise belirli bir tarih aralığında yazılmış tivit metinleri üzerinden kullanılan ilaçların ve yan etkilerinin analiz edilmesi üzerinedir. Nguyen ve diğ. [7], insan algısında sosyal medyanın etkisi üzerine bir çalışma gerçekleştirmişlerdir. Claster ve diğ. [8] çalışmalarında turistlerin bir belge hakkındaki yazmış oldukları tivit metinleri üzerinden algı analizi gerçekleştirmişlerdir. Liu ve diğ. [9] yaptıkları çalışmada, bir şirket hakkında bloglarda yazılmış metinler üzerinden şirketin satış grafiğini çıkarmaya çalışmışlardır Asur ve Huberman [10] çalışmalarında üç milyon tivit üzerinde, Joshi ve diğ. [11] ise film eleştirilerini kullanarak vizyondaki filmler hakkında tahminler çıkarmaya çalışmışladır. Bollen ve diğ. [12] çalışmalarında Twitter verilerini kullanarak borsa hareketlerini tahmin etmeye çalışmışlardır.

Danışmanlı öğrenme kullanılarak gerçekleştirilmiş duygu analizi çalışmaları; Movie Review veri kümesi kullanılarak yapılan çalışmalar; Pang ve diğ. [13] tarafından gerçekleştirilen çalışmada, film yorumlarını sınıflandırmak için 3 farklı makine öğrenmesi algoritması kullanımışlar ve en yüksek doğruluk değerine %82.9 ile destek vektör makineleri ulaşmıştır. Pang ve Lee [14] gerçekleştirdikleri çalışmada bir metin içerisindeki ifadelerin nesnel veya öznel şeklinde sınıflandırılması üzerinedir ve çalışma öznel tümcelerin daha ayırt edici olduğu yargısına varmışlardır. Whitelaw ve diğ. [15] çalışmalarında metinler arasındaki anlamsal ilişkilerin belirlenmesi ve destek vektör makineleri algoritmasıyla %90.2'lik doğru sınıflandırma başarısına ulaşılmışlardır. Yessenalina ve diğ. [16] görüş çıkarımı için belge düzeyinde çok katmanlı bir mimari önermişler ve makine öğrenmesi algoritmaları ile bu görüşleri sınıflandırmaya çalışmışlardır. Matsumoto ve diğ. [17], yapılan değerlendirme belge düzeyinde cümlelerin sınıflandırması üzerine bir yöntem önerisinde bulunmuşlar ve çalışma sonucunda %93.7'lik doğru sınıflandırma doğru sınıflandırma başarısına ulaşmışlardır. Tan ve Zhang [18], görüş madenciliği üzerine yaptıkları çalışmada, Çince metinler üzerinde çalışmışlar ve makine öğrenmesi algoritmaları ile sonuçları sınıflandırmışlardır. Qui ve diğ. [19], görüş sınıflandırmayı reklam sektörü için uygulamışlardır. Bai [20], iki aşamalı bir yöntem önerisi ile kullanıcı yorumlarını makine öğrenmesi algoritmaları ile sınıflandırmışlardır. Chen ve Tseng [21], iki farklı veri kümesi üzerinde yorumları niteliklerine göre sınıflandırmaya çalışmışlardır. Xia ve diğ. [22], kollektif sınıflandırıcı tabanlı bir yorum sınıflandırıcı çalışması gerçekleştirmişlerdir. Kang ve diğ. [23] bir yemek firması için yapılan kullanıcı yorumlarını değerlendirmek için Naïve Bayes yöntemi üzerinde iyileştirme çalışmaları yapmışlardır. Li ve Li [24], sosyal medya siteleri üzerinde yazılan metinlerin özetini sunan bir yapı geliştirmişlerdir. Moraes ve diğ. [25], makine öğrenmesi algoritmalarını kullanarak döküman seviyesinde sınıflandırma yapmışlardır. Wang ve diğ. [26], on veri kümesi üzerinde farklı sınıflandırıcılar kullanarak fikir madenciliği üzerine bir deneysel çalışma

gerçekleştirmişlerdir. Chalothom ve Ellman [27], tivit metinleri üzerinde makine öğrenmesi algoritmaları ile tekil ve topluluk sınıflarıcılar kullanarak sınıflandırma çalışması yapmışlardır. Zheng ve diğ. [28], öznitelik seçmenin önemini ve sınıflandırıcıların performansına etkisini fikir madenciliği konusunda araştırmışlardır.

Yarı Danışmanlı öğrenme kullanılarak gerçekleştirilmiş duygu analizi çalışmaları; Aue ve Gamon [29], etiketli verinin az olduğu durumlarda eğitim için farklı yöntemler önererek, bu yöntemlerin başarıları araştırılmıştır. Tan ve diğ. [30], sıklıkla karşılaşılan bir problem türü olan alan-transfer (domain-transfer)'in çözümü için yeni yöntem önermişler ve bu yöntemin başarımını ölçmüşlerdir. Blitzer ve diğ. [31], alan-transfer probleminin çözümünde yapısal yazışma öğrenme (structural correspondence learning) tabanlı bir yöntem önermişler ve başarımını test etmişlerdir. Mihalcea ve diğ. [32], İngilizce için geliştirilmiş görüş sözlüğünün farklı diller üzerindeki performansını ölçümlemişlerdir. Li and Zong [33], çok-alanlı duygu sınıflandırma (multi-domain sentiment classification) için yeni iki farklı metod denemişlerdir. Banea ve diğ. [34], makine çevirisi (machine translation) tabanlı önerdikleri yeni yöntem ile İngilizce için kullanılan sözlüklerin Romanca ve İspanyolca üzerindeki etkilerini araştırmışlardır. Dasgupta ve Ng [35], yarı-danışmanlı bir yöntem önermişlerdir. Bu yöntem aktif öğrenme (active learning), transdüktif öğrenme (transductive learning) ve spektral kümeleme (spectral clustering) metodları temellidir. Wan ve diğ. [36], eş-eğitim (co-training) temelli Çince yorumların duygusal sınıflandırmasını yapmak için İngilizce kaynaklardan yararlanan bir sistem geliştirmişlerdir. He ve Zhou [37], sözlük tabanlı etiketleminin maliyetlerini düşürecek kendi kendine eğitim (self-training) metodolojine dayanan yeni bir yöntem önermişlerdir. Hernandez ve diğ. [38], duygusal sınıflandırma da kullanılmak üzere birbiriyle ilişkili üç hedef değişkenini (öznellik, görüş kutbu ve etkileme durumu) saptayabilmek için çalışmalar gerçekleştirmişlerdir. Hajmohammadi [39], aktif öğrenmeyi (active learning) yarı danışmanlı olarak kullanarak duygu sınıflandırmak için kullanmışlardır. Hajmohammadi [40], duygu madenciliği için, Hajmohammadi [39]'da sunulan yöntemleri farklı dillerde uygulamışlar ve yarı-danışmanlı eğitimle birleştirerek yeni bir model önerisinde bulunmuşlardır. Hajmohammadi ve diğ. [41], yarı-danışmanlı öğrenme ile kendi kendine eğitim metodlarını birleştirerek ingilizce için oluşturulan derlemlerin diğer diller içinde kullanılmasına imkan sağlamaya çalışmışlardır.

Twitter üzerindeki Türkçe tivitler için yapılan duygu analizi çok fazla değildir. Eroğul [42], bir haber tiviti üzerindeki duyguyu olumlu ve olumsuz olarak sınıflandırma üzerine bir çalışma gerçekleştirmiştir. Vural [43], sözlük temelli bir yaklaşım ile film kritiklerini iki sınıflı (olumlu, olumsuz) şeklinde sınıflandırmıştır. Meral ve Diri [44], Twitter üzerinden toplanan metinler üzerinde, metin temsil yöntemi olarak n-gramları kullanarak üç sınıflı (olumlu-olumsuz-nötr) bir sınıflandırma işlemi gerçekleştirmişlerdir. Şimşek ve Özdemir [45], atılan tivitler ile hisse senetlerinin değişimlerini arasındaki ilişkiyi modelleyecek bir sistem üzerine çalışmışlardır. Türkmenoğlu ve Tantuğ [46], film kritikleri ve tivitler üzerinden sözlük tabanlı ve makine öğrenmesi algoritmalarını kullanarak duygu çıkarımı yapmışlardır.

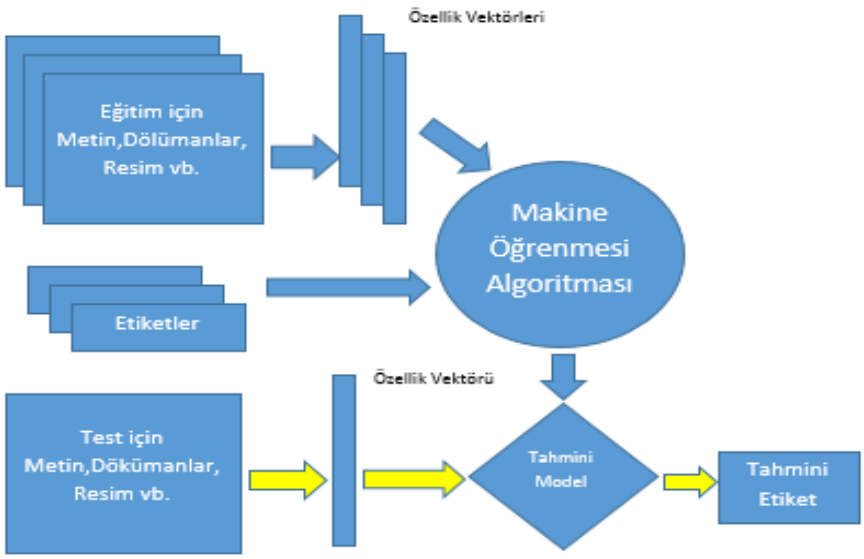
2. Deneysel metod

Makine Öğrenmesinde kullanılan öğrenme yöntemlerinden Danışmanlı Öğrenme-Yarı Danışmanlı Öğrenme yöntemleri, döküman sınıflandırma için kullanılan Doc2Vec yöntemi ve kullanılan veri kümesi ile ilgili bilgiler verilecektir.

2.1. Danışmanlı öğrenme

Bu tip öğrenmede öğrenen sistemin olayı öğrenmesine bir danışman yardımcı olur. Danışman sistemin öğrenmesi istenen konuyla ilgili örnekleri Girdi / Çıktı olarak verir. Danışmanlı öğrenme, girişlerle çıkışların eşleştiği örneklerden bir fonksiyonun öğrenilmesi ya da hipotezin bulunmasıdır [47]. Bu tip öğrenmede mutlaka bir danışmana ihtiyaç vardır. Danışmanlı öğrenme ait gösterim Şekil 1'de verilmiştir.

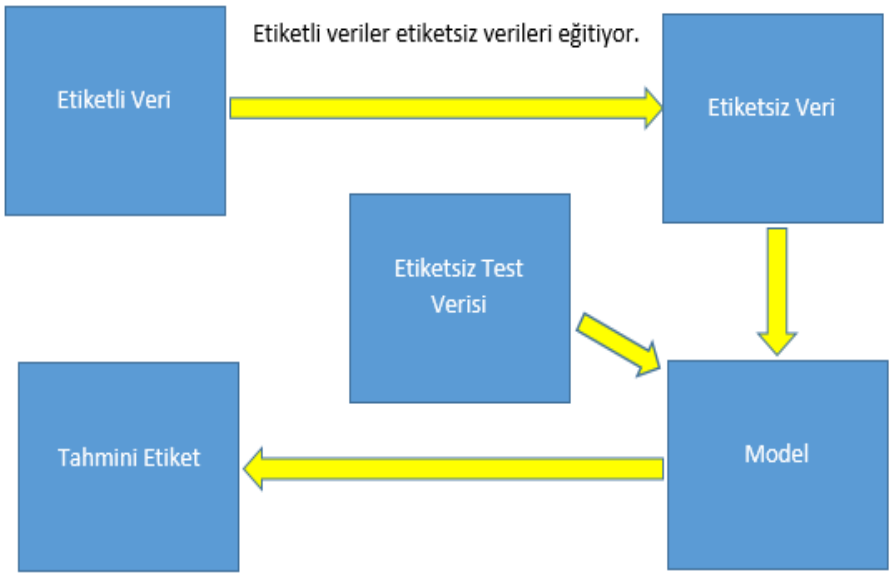
Danışmanlı öğrenme modelinde sınıflandırma işlemi gerçekleştirilir ve eğitim aşamasında oluşturulan model ile test aşamasında sistemin sınıflandırma yapması beklenir [48]. Yaygın olarak kullanılan algoritmalar, Destek Vektör Makinesi (Support Vector Machine), Yapay Sinir Ağları (Artificial Neural Network), Naive Bayes, k-En Yakın Komşu (k-Nearest Neighbour) ve Karar Ağaçları (Decision Trees) 'dır [49]. Sınıflandırma işlemi sonucunda tek etiketli (single-label) veya çok etiketli (multi-label) sınıflandırma gerçekleştirilebilir [50].



Şekil 1. Danışmanlı Öğrenme (Supervised Learning) [51].

2.2. Yarı danışmanlı öğrenme

Makine öğrenmesinde kullanılan öğrenme yöntemlerinden biridir. Giriş olarak verilen veri de çok miktarda etiketlenmemiş veri ve az miktarda etiketlenmiş veri sunulmaktadır [52]. Bu metot genellikle etiketlenmiş verinin az olduğu, etiketlenmemiş verinin kolaylıkla elde edilebildiği durumlarda kullanışlı olmaktadır. Yarı danışmanlı öğrenme ait gösterim Şekil 2'de verilmektedir [51].

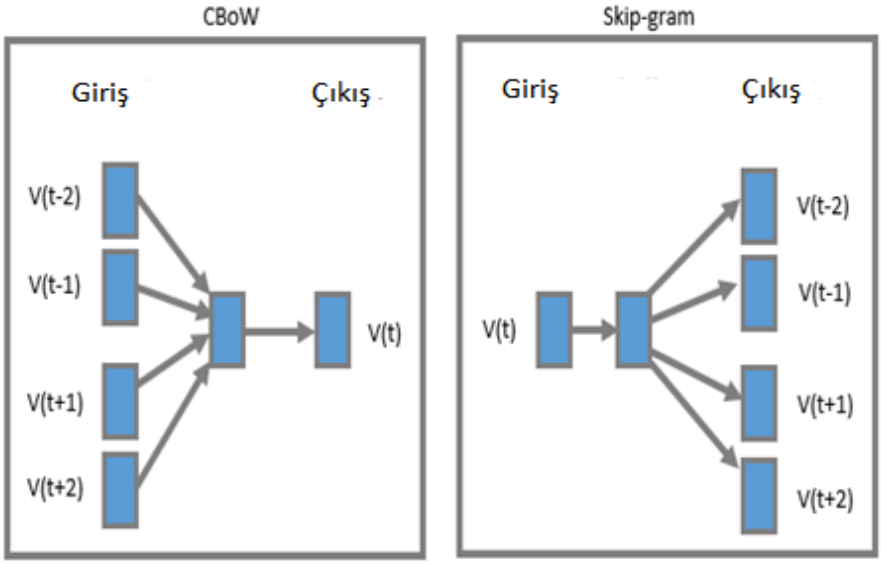


Şekil 2. Yarı Danışmanlı Öğrenme (Semi-Supervised Learning).

2.3. Doc2Vec

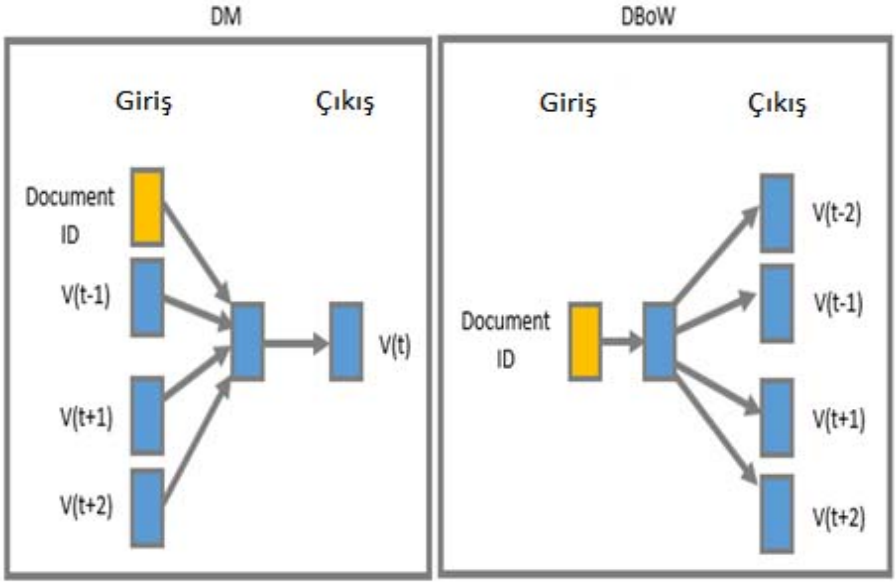
Doc2Vec yönteminin temeli Word2Vec yöntemine dayanmaktadır. Word2Vec algoritması, kelimeleri uzaysal bir düzleme taşıyarak bir vektör oluşturma yöntemidir. Doc2Vec ise cümleler ya da paragraflar için benzer işlemleri gerçekleştirmektedir. Doc2Vec yöntemi bazı kaynaklarda Paragraph2Vec ismiyle de anılmaktadır. Word2Vec algoritmasının dokümanlar için özelleştirilmiş halidir [1,4,5]. Word2Vec algoritmasına göre yapılan değişiklik Şekil 3 ve Şekil 4'ten de görüleceği üzere doküman numarası (document ID) eklenmesidir. Word2Vec'de amaç eğitim sürecinde ortalama olasılığı maksimize etmektir. Eş.1'de Doc2Vec yöntemine ait eşitlik görülmektedir. Eş. 1'de $w_1, w_2, \ldots, w_T$ eğitim kelimelerinin dizilimini vermektedir.

$$\frac{1}{T}\sum_{t=k}^{T-k} \log p\left(w_t | w_{t-k}, \ldots, w_{t+k}\right) \quad (1)$$



Şekil 3. Word2Vec - CBoW ve Skip-gram (Word2Vec-CBOW and Skip-gram) [51].

Word2Vec algoritmasında Continuous bag of words (CBoW) ve Skip-Gram (SG) şeklinde iki farklı yöntemi mevcuttur. Bu yöntemler Doc2Vec için yapılandırılarak Distributed Memory (DM) ve Distributed bag of words (DBoW) adında iki yönteme dönüştürülmüştür.



Şekil 4. Doc2Vec-DM ve DBoW (Doc2Vec- DM and DBoW) [51].

2.4. Veri kümeleri (Data sets)

Yapılan çalışmada Türkçe ve İngilizce için iki farklı veri kümesi kullanılmıştır. Türkçe için yapılan çalışmada içerisinde 2906 tivit içeren duygu analizi için hazırlanmış bir veri kümesi kullanılmıştır [53]. Veri kümesi oluşturulurken Twitter üzerinden özel bir telekom şirketine ait gönderiler seçilmiştir. Veri kümesine ait bilgiler Tablo 1'de görülmektedir. Etiketsiz eğitim kümesi ise Twitter üzerinden yazdığımız api yardımıyla elde edilen tivitlerin düzenlenmesiyle oluşturulmuştur ve 2281 cümlelik bir veri kümesidir.

Tablo 1. Türkçe veri kümesi.

Sınıf	Cümle Sayısı
Etiketli	
Poztif	724
Negatif	1270
Nötr	912
Toplam	2906
Etiketsiz	
2281 Cümle	

İngilizce için yapılan çalışmada içerisinde 1774 adet etiketsiz ve 58817 adet etiketli veri içeren veri kümesi kullanılmıştır [54]. Amerikadaki 6 büyük havayolu şirketinin (American, Delta, Southwest Airlines, United, US Airways, ve Virgin America) müşterilerinin 1 hafta içinde Twitter'da paylaştıkları tivitler toplanarak veri kümesi oluşturulmuştur. Veri kümesine ait bilgiler Tablo 2'de görülmektedir.

Tablo 2. İngilizce veri kümesi.

Sınıf	Cümle Sayısı
Etiketli	
Poztif	301
Negatif	1091
Nötr	382
Toplam	1774
Etiketsiz	
58817 Cümle	

Veri kümeleri üzerinde Doc2Vec algoritmalarının çalıştırılmasından önce veri üzerinde bazı ön işlemler gerçekleştirilmiştir. Ön işlem safhasında duygu analizi için bir önemi olmayan veriler silinmiştir. Böylelikle hem sistemin başarı oranı artırılmış hem de gereksiz bilgiler ile uğraşılması önlenmiş ve böylelikle sistemin hesaplama maliyeti düşürülmüş olmaktadır. Tablo 3'te temizlenen verilere örnekler verilmiştir.

Tablo 3. Önişlem safhasında silinen veriler.

Silinen Karakterler
Html Etiketleri
Twitter Kullanıcı Adları
Twitter Hashtagleri
Telefon numarası vb.
Posta kodları
Düzenli ifadeler

Twitter üzerinden elde edilen tivitler Json formatındadır. Türkçe için özel karakterler olan "ç","ş" vb. için dönüşümler yapılması sınıflandırmanın doğruluğunu artırmak için çok önemlidir. Alınan ham veriler üzerinde dönüşüm işlemleri gerçekletirilmiştir. Gerçekleştirilen çalışma da DM algoritması için kullanılan parametre değerleri Tablo 4'te görülmektedir. DBoW algoritması için dm_concat parametresi yoktur ve dm parametresi 0 (sıfır) değerini almaktadır.

Tablo 4. DM'nin kullandığı parametreler (Parameters used for DM).

Parametre	Değeri
Size	400
Window	8
Min_count	1
Sample	1e-4
negative	5
workers	4
dm	1
dm_concat	1

3. Bulgular ve tartışmalar

Gerçekleştirilen çalışmada iki farklı veri kümesi için iki farklı Doc2Vec algoritması çalıştırılmıştır. Çalışma sonuçlarının değerlendirilmesinde doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F-ölçüsü metrikleri kullanılmıştır.

Doğruluk metriği, sınıflandırma başarısının tespit edilmesinde kullanılan en önemli metriklerdendir. Doğru sınıflandırılmış örnek sayısının (True Positive + True Negative), toplam örnek sayısına (TP + TN + False Pozitive + False Negative) oranıdır [55]. Doğruluk metriğine ait denklem Eş. 2'de görülmektedir.

$$Doğruluk = \frac{(TP)+(TN)}{TP+FP+FN+TN} \tag{2}$$

Kesinlik metriği, sınıfı doğru olarak belirlenmiş örneklerin TP (True Positive), sınıfı doğru olarak belirlenmiş tüm örneklere oranıdır. Kesinlik metriğine ait denklem Eş. 3'te görülmektedir.

$$Kesinlik = \frac{(TP)}{TP+FP} \tag{3}$$

Duyarlılık metriği, doğru sınıflandırılmış pozitif örneklerin TP sayısının, toplam pozitif örneklere oranıdır. Duyarlılık metriğine ait denklem Eş. 4'te görülmektedir.

$$Duyarlılık = \frac{(TP)}{TP+FN} \tag{4}$$

F-ölçütü ise Kesinlik ve Duyarlılık metriğinin harmonik ortalamasıdır. F-ölçütüne ait denklem Eş. 5'te görülmektedir.

$$F-ölçütü = \frac{2*Kesinlik*Duyarlılık}{Kesinlik+Duyarlılık} \tag{5}$$

Yarı danışmanlı öğrenmenin uygulanabilmesi için Python programlama dilinde Gensim kütüphanesi kullanılarak bir yazılım geliştirilmiştir. Geliştirdiğimiz yazılım yarı danışmanlı olarak sistemi eğitebilecek şekilde tasarlanmıştır. Etiketli veriler ile etiketsiz veriler sisteme beraber verilmektedir. Oluşturulan yazılım etiketli ve etiketsiz verileri kullanarak bir model oluşturmaktadır. Ardından oluşturulan bu model yardımıyla test verilerinin etiketlenmesi sağlanmaktadır. Yazılım Doc2Vec yönteminin iki ayrı metodu olan DBoW ve DM için kodlanmıştır. Ayrıca gerçekleştirilen yazılım yardımıyla doğruluk (accuracy) değerleri hesaplanmış ve karışım (confusion) matrisleri oluşturulmuştur. Danışmanlı öğrenme için ise Weka yazılımı kullanılmıştır. Makine öğrenmesi algoritması olarak da Destek Vektör Makineleri (Support Vector Machines) kullanılmıştır. Tablo 1'de verilen Türkçe veri kümesi için yapılan çalışmaya ait sonuçlar Tablo 5'te görülmektedir.

Tablo 5. Türkçe veri kümesi için sonuçlar.

Yöntem	Cümle Sayıları						
Yarı Danışmanlı	250	500	750	1000	1500	2000	2500
DM	0.416	0.448	0.4413	0.443	0.4306	0.4315	0.4328
DBoW	**0.44**	**0.452**	**0.4506**	**0.46**	**0.448**	**0.4405**	**0.4488**
Danışmanlı							
DM	0.368	0.364	0.3706	0.396	0.3746	0.358	0.3512
DBoW	**0.432**	**0.4528**	**0.4502**	**0.441**	**0.4393**	**0.432**	**0.4252**

Tablo 6. Türkçe veri kümesi için karışım matrisleri (Cümle Sayısı=2500).

Yarı Danışmanlı Öğrenme	DM	Negatif	Nötr	Pozitif
	Negatif	**1037**	35	16
	Nötr	728	**18**	35
	Pozitif	577	24	**25**
	DBoW	Negatif	Nötr	Pozitif
	Negatif	**1058**	20	10
	Nötr	723	**30**	28
	Pozitif	575	22	**34**
Danışmanlı Öğrenme	DM	Negatif	Nötr	Pozitif
	Negatif	**437**	333	263
	Nötr	314	**226**	284
	Pozitif	242	186	**215**
	DBoW	Negatif	Nötr	Pozitif
	Negatif	**738**	252	43
	Nötr	520	**247**	57
	Pozitif	385	180	**78**

Tablo 2'de verilen İngilizce veri kümesi için yapılan çalışmaya ait sonuçlar Tablo 7'te görülmektedir.

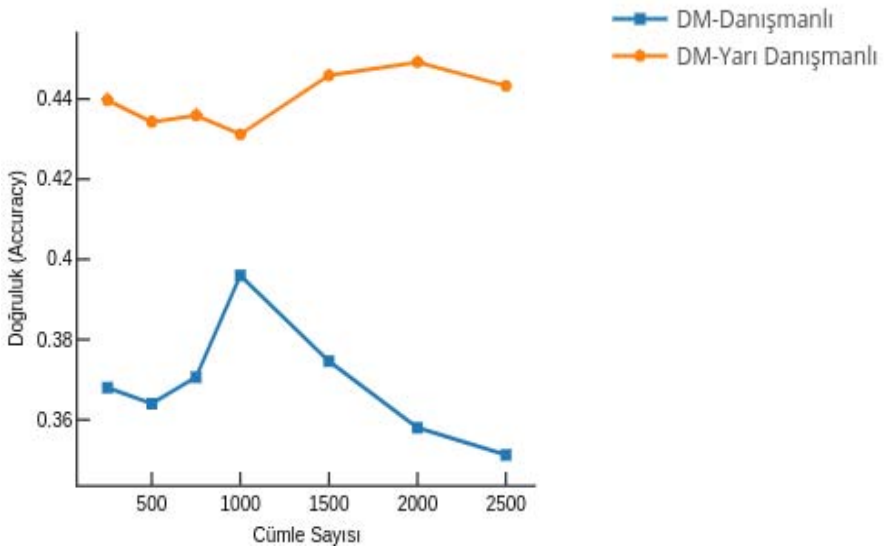
Tablo 7. İngilizce için sonuçlar (Results for English).

Yöntem	Cümle Sayıları					
Yarı Danışmanlı	250	500	750	1000	1250	1500
DM	0.604	0.626	0.620	0.601	0.6128	0.6066
DBoW	**0.608**	**0.632**	**0.635**	**0.610**	**0.6216**	**0.6206**
Danışmanlı						
DM	0.58	0.6006	0.5733	0.552	0.5376	0.518
DBoW	**0.5902**	**0.612**	**0.6013**	**0.602**	**0.6056**	**0.606**

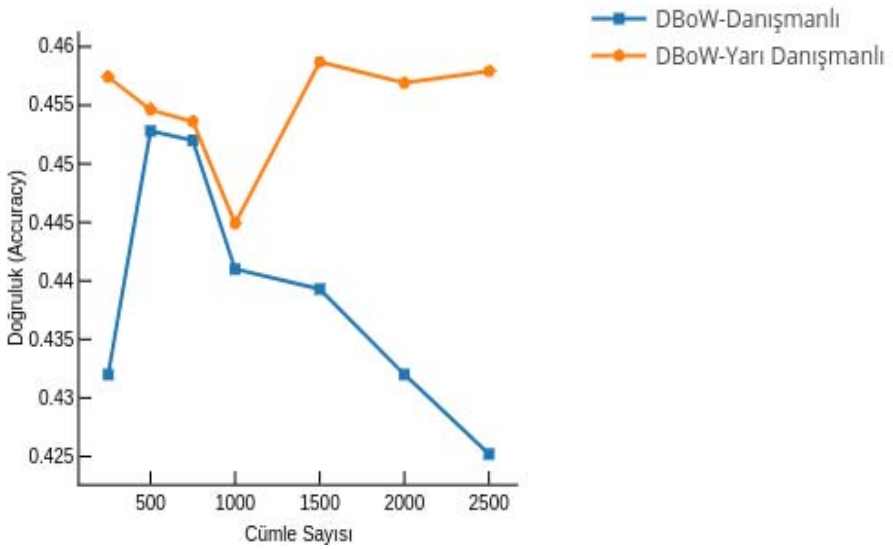
Tablo 8. İngilizce Veri Kümesi için Karışım Matrisleri (Cümle sayısı=1500).

Yarı Danışmanlı Öğrenme	DM	Negatif	Nötr	Pozitif
	Negatif	**861**	50	10
	Nötr	270	**43**	7
	Pozitif	229	25	**5**
	DBoW	Negatif	Nötr	Pozitif
	Negatif	**873**	43	5
	Nötr	265	**49**	6
	Pozitif	225	25	**8**
Danışmanlı Öğrenme	DM	Negatif	Nötr	Pozitif
	Negatif	**710**	108	91
	Nötr	253	**44**	40
	Pozitif	188	43	**23**
	DBoW	Negatif	Nötr	Pozitif
	Negatif	**801**	52	56
	Nötr	265	**69**	33
	Pozitif	175	39	**40**

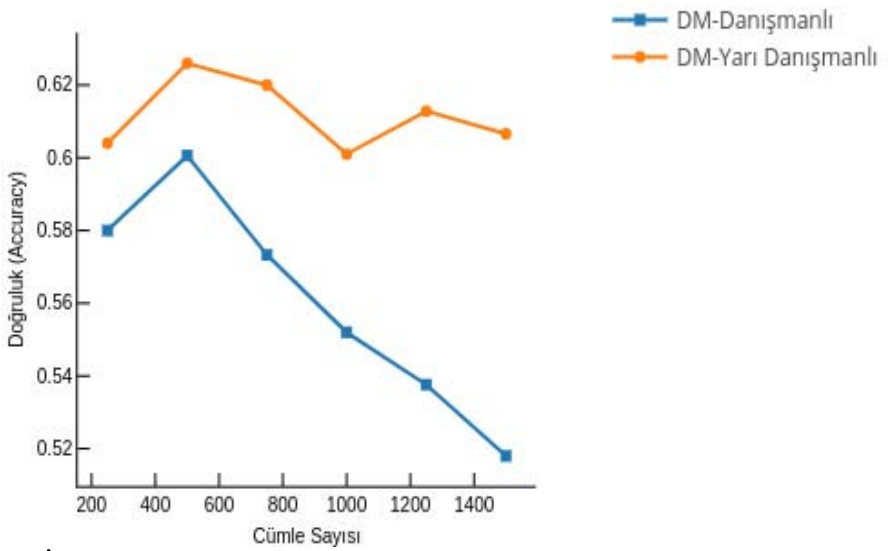
Türkçe ve İngilizce veri kümeleri yarı danışmanlı öğrenme ve danışmanlı öğrenme uygulanarak çalıştırılmıştır. Elde edilen doğruluk metriğine ait grafikler Şekil 5-8'de görülmektedir.



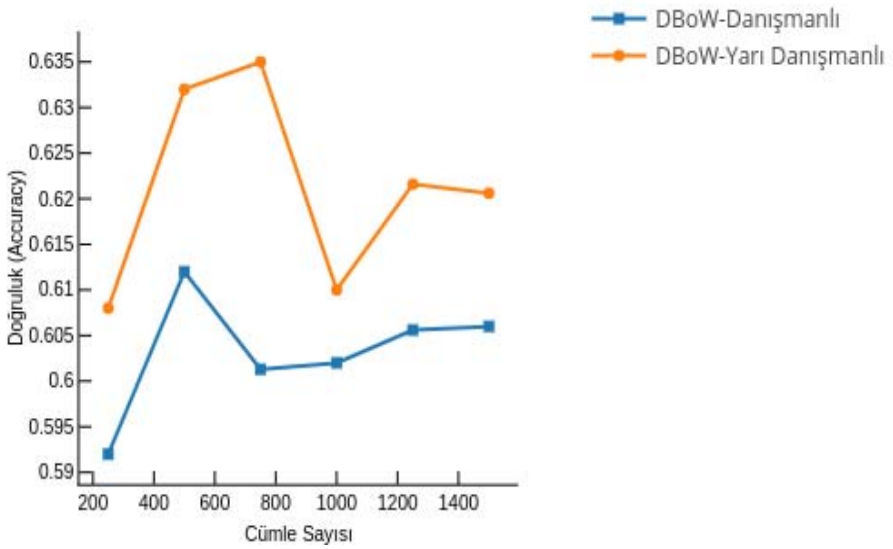
Şekil 5. Türkçe veri kümesine için DM yönteminin sonuçları.



Şekil 6. Türkçe veri kümesine için DBoW yönteminin sonuçları.



Şekil 7. İngilizce veri kümesine için DM yönteminin sonuçları.



Şekil 8. İngilizce veri kümesine için DBoW yönteminin sonuçları.

Şekil 5-8'te görüldüğü üzere her iki dil içinde Yarı Danışmanlı öğrenme daha yüksek doğruluk değerine ulaşmıştır. DBoW yöntemi de DM'ye göre daha yüksek doğruluk değerine ulaşmıştır.

Doğruluk metriği dışındaki kesinlik, duyarlılık (recall) ve F-ölçüsü metriklerine ait sonuçlar Tablo 9'da görülmektedir.

Tablo 9. Diğer metriklerin sonuçları.

Veri Kümesi	Yöntem	Öğrenme Türü	Kesinlik	Duyarlılık	F-Ölçütü
Türkçe	DM	Yarı Danışmanlı	0,348	0,432	0,3855
	DBoW		**0,4447**	**0,4488**	**0,4467**
	DM	Danışmanlı	0,3586	0,3512	0,3548
	DBoW		**0,4182**	**0,4252**	**0,4216**
İngilizce	DM	Yarı Danışmanlı	0,5062	0,6067	0,5519
	DBoW		**0,5556**	**0,6206**	**0,5863**
	DM	Danışmanlı	0,4497	0,518	0,481
	DBoW		**0,5502**	**0,6066**	**0,577**

XX

Hem Türkçe, hemde İngilizce veri kümesi için kesinlik, duyarlılık ve F-ölçütü değerleri en yüksek değerlerine yarı danışmanlı öğrenme ve DBoW yöntemiyle ulaşmıştır.

4.Sonuçlar

Bu çalışmanın ilk amacı Yarı Danışmanlı ve Danışmanlı eğitimin elimizdeki veri kümelerinin doğru sınıflandırılması üzerindeki etkilerinin araştırılmasıdır. İkinci amacımız ise döküman vektörü olarak isimlendirilen Doc2Vec algoritmasının iki farklı metodu olan DBoW ve DM yöntemlerinin aynı şartlar altındaki performanslarının araştırılmasıdır.

Hem Türkçe hemde İngilizce dilleri için hem etiketli hemde etiketsiz veri kümelerinin oluşturulması ile işe başlanmıştır. Elde edilen veri kümeleri üzerinde çeşitli ön işlemler gerçekleştirildikten sonra veriler eğitim için hazır hale getirilmiştir. Eğitim aşamasında elimizdeki veriler yarı danışmanlı ve danışmanlı öğrenme kullanılarak Doc2Vec algoritmasının iki farklı yöntemi DBoW ve DM üzerinde çalıştırılarak bir model oluşturulmuştur. Oluşturulan modeller kullanılarak test aşamasına geçilmiştir. Çalışma sonuçları 4 farklı metrik (doğruluk, kesinlik, duyarlılık, F-ölçütü) cinsinden verilmiştir.

Gerçekleştirilen çalışmanın sonuçlarından ilki yarı danışmanlı eğitimin danışmanlı eğitime kıyasla daha yüksek doğruluk, kesinlik, duyarlılık ve f-ölçütü değerlerine ulaşmasıdır. Dolayısıyla makine öğrenmesi yöntemleri ile bir eğitim gerçekleştirirken etiketli veriyle beraber etiketsiz verilerinde kullanılmasının oluşturulan modelin doğruluğunu artırdığı söylenebilir.

Çalışma sonunda elde ettiğimiz ikinci yararlı sonuç ise Doc2Vec algoritmasının iki farklı yöntemi arasındaki sınıflandırma başarılarının farklılığıdır. Ölçülen 4 farklı metrik değeri için hem yarı danışmanlı hemde danışmanlı öğrenme yöntemi kullanıldığında DBoW yöntemi DM yöntemine göre daha yüksek başarı elde etmiştir. DM, DBoW yöntemine göre daha fazla veriye ihtiyaç duyması ve elimizdeki veri kümelerindeki örnek sayısının az olması DM yönteminin DBoW yönteminden geri kalmasının nedenlerinden birisi olabilir.

Gerçekleştirilen çalışma Doc2Vec yönteminin Türkçe için duygu analizi için gerçekleştirilmiş ilk çalışmalarından olması sebebiyle önemlidir. Türkçe için alınan sonuçların İngilizce'den daha düşük olmasının nedeni Türkçe için kullanılan veri kümesinin daha küçük olması gösterilebilir.

Bundan sonraki çalışmalarımızda Türkçe etiketli veri kümesinin büyütülmesi ve veri kümesindeki etiketli veri sayısının sistemin başarısı üzerindeki etkilerinin araştırılması öncelikli hedefimizdir. Böylece DM ve DBoW yöntemleri arasındaki veri miktarı arasındaki ilişkinin de ölçülmesi planlanmaktadır. Ayrıca danışmansız öğrenmenin elimizdeki veriler için elde edeceği sonuçlarda geleceğe dair yapılması planlanan görevlerden biridir. Bir başka gelecek planımız ise DBoW ve DM'nin hibrit bir yaklaşım şeklinde duygu analizinde kullanılabilirliğinin araştırılması ve sonuçlarının elde edilmesi üzerinedir.

Kaynaklar

[1] Go, A., Huang, L., and Bhayani, R., Twitter sentiment analysis, **Entropy**, 17, (2009).

[2] Liu, B. and Lei, Z., Mining Text Data: A survey of opinion mining and sentiment analysis, Mining Text Data, Springer, USA, pp. 415-463. ISBN: 978-1-4614-3223-4.Prabowo, R. and Thelwall, M., "Sentiment analysis: A combined approach", **Journal of Informetrics**, 3(2), 143-157, (2009).

[3] Akgül, E.S., Ertano, C. ve Diri, B., Twitter verileri ile duygu analizi, **Pamukkale University Journal of Engineering Sciences**, 22(2), 106-110, (2016).

[4] Szomszor, M.N., Kostkova, P. and Quincey, E.D., # Swineflu: Twitter predicts swine flu outbreak in 2009, **3rd International ICST Conference on Electronic Healthcare for the 21st Century**. 2012.

[5] Bian J, Topaloglu, U, Yu, F., Towards large-scale Twitter mining for drug-related adverse events, **International Workshop on Smart Health and Wellbeing (SHB'12)**, Maui, Hawaii, USA, 29 October-2 November 2012.

[6] Nguyen, L.E., Wu, P., Chan, W., Peng, W., Zhang, Y., Predicting collective sentiment dynamics from time-series social media, **Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM '12)**, Beijing, China, 12 August 2012.

[7] Claster, W.B., Dinh, H., Cooper, M., Naive bayes and unsupervised artificial neural nets for Cancun tourismsocial media data analysis, **2nd World Congress on Nature and Biologically Inspired Computing (NaBIC)**. Kitakyushu, Fukuoka, Japan, 15-17 December 2010.

[8] Liu, Y., Huang, X., An, A., Yu, X., ARSA: A sentiment awaremodel for predicting sales performance using blogs, **30th ACM SIGIR International Conference on Research and Development in Information Retrieval**, Amsterdam, the Netherlands, 23-27 July 2007.

[9] Asur, S., Huberman, B.A., Predicting the Future with Social Media, **IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT)**, Toronto, ON, Canada, 31 August-3 September 2010.

[10] Joshi, M., Das, D., Gimpel, K., Smith, N.A., Movie reviews and revenues: an experiment in text regression, **Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)**, Los Angeles, CA, USA, 1-6 June 2010.

[11] Bollen, J., Mao, H., Zeng, X., Twitter mood predicts the stock market, **Journal of Computational Science**, 2(1), 1-8, (2011).

[12] Pang, B., Lee, L. and Vaithyanathan S., Thumbs up?: sentiment classification using machine learning techniques, Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. **Association for Computational Linguistics**, 2002.

[13] Pang, B., and Lillian, L., A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, Proceedings of the 42nd annual meeting on Association for Computational Linguistics. **Association for Computational Linguistics**, 2004.

[14] Whitelaw, C., Garg, N., Argamon, S., Using appraisal groups for sentiment analysis, **14th ACM International Conference on Information and**

Knowledge Management (CIKM), Bremen, Germany, 31 October-5 November 2005.

[15] Yassenalina, A., Yue, Y., Cardie, C., Multi-Level structured models for document-level sentiment classification, **Conference on Empirical Methods in Natural Language Processing (EMNLP)**, Boston, MA, USA, 9-11 October 2010.

[16] Matsumoto, S., Takamura, H., Okumura, M., Sentiment classification using word sub-sequences and dependency sub-trees, **9th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)**, Hanoi, Vietnam, 18-20 May 2005.

[17] Tan, S., Zhang, J., An empirical study of sentiment analysis for Chinese document, **Expert Systems with Applications**, 34(4), 2622-2629, (2008).

[18] Qui, G., He, X., Zhang, F., Shi, Y., Bu, J., Chen, C., DASA: dissatisfaction-oriented advertising based on sentiment analysis, **Expert Systems with Application**, 37(9), 6182-6191, (2010).

[19] Bai, X., Predicting consumer sentiments from online text, **Decision Support Systems**, 50(4), 732-742, (2011).

[20] Chen, C.C., Tseng, Y.D., Quality evaluation of product reviews using an information quality framework, **Decision Support Systems**, 50(4), 755-768, (2011).

[21] Xia, R., Zong, C., Li, S., Ensemble of feature sets and classification algorithms, **Information Sciences**, 181(6), 1138-1152, (2011).

[22] Kang, H., Yoo, S.J., Han, M., Senti-Lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews, **Expert Systems with Applications**, 39(5), 6000-6010, (2012).

[23] Li, Y.M., Li, T.Y., Deriving Market intelligence from microblogs, **Decision Support Systems**, 55(1), 206-217, (2013).

[24] Moraes, R., Valiati, J.F., Neto WPG, Document-Level sentiment classification: an empirical comparison between SVM and ANN, **Expert Systems with Applications**, 40(2), 621-633, (2013).

[25] Wang, G., Sun, J., Ma, J., Xu, K., Gu, J., Sentiment classification: the contribution of ensemble learning, **Decision Support Systems**, 57, 77-93, (2014).

[26] Chalothom, T., Ellman, J., Simple Approaches of Sentiment Analysis via Ensemble Learning, Editor: Kim KJ. **Information Science and Applications**, 631-639, Berlin, Germany, Springer, 2015.

[27] Zheng, L., Wang, H., Gao, S., Sentimental feature selection for sentiment analysis of Chinese online reviews, **International Journal of Machine Learning and Cybernetics**, 1-10, (2015).

[28] Aue, A., Gamon, M., Customizing sentiment classifiers to new domains: a case study, **International Conference on Recent Advances in Natural Language Processing (RANLP)**, Borovets, Bulgaria, 21-23 September 2005.

[29] Tan, S., Wu, G., Tang, H., Cheng, X., A novel scheme for domain-transfer problem in the context of sentiment analysis, **Conference on Information and Knowledge Management (CIKM)**, Lisbon, Portugal, 6-10 November 2007.

[30] Blitzer, J., Dredze, M., Pereira, F., Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification, **45th Annual Meeting of the Association for Computational Linguistics (ACL)**, Prague, Czech Republic, 25-27 June 2007.

[31] Mihalcea, R., Banae, C., Wiebe, J., Learning multilingual subjective language via cross-lingual projections, **45th Annual Meeting of the Association for Computational Linguistics (ACL)**, Prague, Czech Republic, 25-27 June 2007.

[32] Li, S., Zong, C., Multi-Domain sentiment classification, **46th Annual Meeting of the Association for Computational Linguistics (ACL)**, Columbus, OH, USA, 19-20 June 2008.

[33] Banae, C., Mihalcea, R., Wiebe, J., Multilingual subjectivity analysis using machine translation, **Conference on Empirical Methods in Natural Language Processing (EMNLP)**, Honolulu, HI, USA, 25-27 October 2008.

[34] Dasgupta, S., Ng, V., Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification, **47th Annual Meeting of the Association for Computational Linguistics (ACL)**, Suntec, Singapore, 2-7 August 2009.

[35] Wan, X., Co-training for cross-lingual sentiment classification, **47th Annual Meeting of the Association for Computational Linguistics (ACL)**, Suntec, Singapore, 2-7 August 2009.

[36] He, Y., Zhou, D., Self-Training from labelled features for sentiment analysis, **Information Processing and Management**, 47(4), 606-616, (2011).

[37] Hernandez, O.J., Rodriguez, J.D., Alzate, L., Lucania, M., Inza, I., Lozano, J.A., Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers, **Neurocomputing**, 92, 98-115, (2012).

[38] Hajmohammadi, M.S., Ibrahim, R., Selamat, A., Bi-View semi-supervised active learning for cross-lingual sentiment classification, **Information Processing and Management**, 50(5), 718-732, (2014).

[39] Hajmohammadi, M.S., Ibrahim, R., Selamat, A., Cross-Lingual sentiment classification using multiple source languages in multi-view semi-supervised learning, **Engineering Applications of Artificial Intelligence**, 36, 195-203, (2014).

[40] Hajmohammadi, M.S., Ibrahim, R., Selamat, A., Fujita, H., Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples, **Information Sciences**, 317, 67-77, (2015).

[41] Eroğul, U., Sentiment Analysis in Turkish. **MSc Thesis, Middle East Technical University**, Ankara, Turkey, 2009.

[42] Vural, A.G., Cambazoğlu BB, Şenkul P, Tokgöz ZO., A frame work for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish, **27th International Symposium on Computer and Information Sciences**, Paris, France, 3-4 October 2012.

[43] Meral, M., Diri, B., Twitter üzerinde duygu analizi, **IEEE 22. Sinyal İşleme ve İletişim Uygulamaları Kurultayı**, Trabzon, Türkiye, 23-25 Nisan 2014.

[44] Şimşek, M., Özdemir, S., Analysis of the relation between Turkish twitter messages and stock market index, **6th International Conference on Application of Information and Communication Technologies (AICT)**, Tbilisi, Georgia, 17- 19 October 2012.

[45] Türkmenoğlu, C., Tantuğ, A.C., Sentiment analysis in Turkish media, **Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM '14)**, Beijing, China, 21-26 June 2014.

[46] Nilsson, N.J., Introduction to machine learning: An early draft of a proposed textbook, (1996).

[47] Chao, W-L., Machine Learning Tutorial, **Disp. Ee. Ntu. Edu. Tw**, (2011).

[48] Caruana, R., and Niculescu-Mizil, A., An empirical comparison of supervised learning algorithms, **Proceedings of the 23rd international conference on Machine learning**. ACM, 2006.
[49] Sebastiani, F., Machine learning in automated text categorization, **ACM computing surveys (CSUR)**, 34.1, 2002.
[50] Bilgin, M., Makine Öğrenmesi,. **Papatya Yayincilik**, Istanbul, (2018).
[51] Witten, I.H., et al., Data Mining: Practical machine learning tools and techniques, **Morgan Kaufmann**, 2016.
[52] Çetin, M., and Amasyalı, M.F., Supervised and Traditional Term Weighting Methods for Sentiment Analysis., **Sinyal İşleme Kurultayı**, (2013)
[53] Airline Twitter Sentiment, https://www.crowdflower.com/data-for-everyone/, Online: April 2017.
[54] Kesim, M., Real time measurement of micro changes in dinamic images, **Msc. Thesis, Karadeniz Technical University**, Trabzon, Turkey, (2015).