Pamukkale Univ Muh Bilim Derg, 22(2), 111-122, 2016

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Makine öğrenmesi yöntemlerinin görüş madenciliğinde kullanılması üzerine bir literatür araştırması

A review of literature on the use of machine learning methods for opinion mining

Aytuğ ONAN[1], Serdar KORUKOĞLU[2]*

[1]Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, Celal Bayar Üniversitesi, Manisa, Türkiye.
aytug.onan@cbu.edu.tr
[2]Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, Ege Üniversitesi, İzmir, Türkiye.
serdar.korukoglu@ege.edu.tr

Geliş Tarihi/Received: 21.01.2015, Kabul Tarihi/Accepted: 20.04.2015 doi: 10.5505/pajes.2015.90018
* Yazışılan yazar/Corresponding author Derleme Makalesi/Review Article

Öz

Görüş madenciliği, görüş sahibinin tutum, davranış, duygu gibi öznel bilgilerinin çıkarılması için doğal dil işleme, metin madenciliği, hesaplamalı dilbilim gibi bilim alanlarının tekniklerini kullanan güncel bir araştırma alanıdır. Görüş madenciliği işleminin temel olarak bir sınıflandırma problemi olarak ele alınması mümkündür. Bu nedenle, makine öğrenmesine dayalı yöntemler sıklıkla görüş sınıflandırma amacıyla uygulanmaktadır. Görüş madenciliğinde makine öğrenmesine dayalı yöntemler temel olarak, öğreticili, yarı-öğreticili ve öğreticisiz yöntemler olmak üzere üç temel sınıf altında incelenmektedir. Bu çalışma kapsamında, görüş madenciliği alanında gerçekleştirilen temel makine öğrenmesine dayalı çalışmalar ve her bir makine öğrenmesi yönteminin güçlü ve zayıf yönleri ele alınmaktadır.

Anahtar kelimeler: Görüş madenciliği, Makine öğrenmesi, Duygu analizi

Abstract

Opinion mining is an emerging field which uses methods of natural language processing, text mining and computational linguistics to extract subjective information of opinion holders. Opinion mining can be viewed as a classification problem. Hence, machine learning based methods are widely employed for sentiment classification. Machine learning based methods in opinion mining can be mainly classified as supervised, semi-supervised and unsupervised methods. In this study, main existing literature on the use of machine learning methods for opinion mining has been presented. Besides, the weak and strong characteristics of machine learning methods have been discussed.

Keywords: Opinion mining, Machine learning, Sentiment analysis

1 Giriş

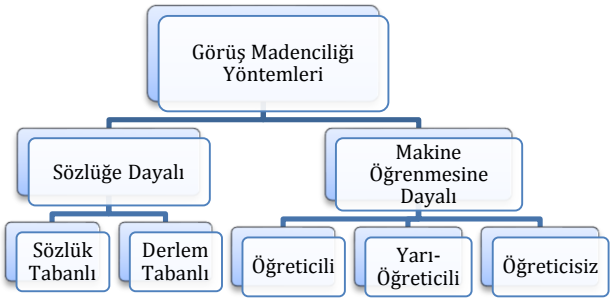
Görüş Madenciliği/Duygu Analizi (Opinion Mining/Sentiment Analysis), bilgisayar bilimleri yöntemleri kullanılarak, görüş sahibinin metin içerisinde belirttiği görüş, duygu ve tutum gibi öznel bilgilerin çıkarılmasını amaçlayan bir araştırma alanıdır [1]. Bilgi ve iletişim teknolojilerindeki ilerlemeler ile birlikte web, diğer insanların belirli bir ürün ya da konu hakkındaki görüşlerine erişilebilen önemli bir bilgi kaynağı durumuna gelmiştir. Web'in e-ticaret sitesi, değerlendirme siteleri, tartışma forumları gibi farklı mecralarında yer alan görüş içeren bilgi hacmindeki artış ile birlikte görüş alışverişinin şekli değişmiştir. Görüş madenciliği işletmelerin, yeni piyasaya sürdükleri bir ürün/hizmet hakkında etkin bir bilgi toplama aracı olarak kullanılabilir. Belirli bir pazara ilişkin bilgi toplamada ya da farklı ürün ve hizmetlerin mevcut ürün/hizmete karşı iyi ve kötü özelliklerinin belirlenmesinde kullanılabilir. Ürün yorumlarına ilişkin reklam yerleştirme de görüş madenciliği kullanım alanları arasındadır. Bireysel düzeyde görüş madenciliği, belirli bir ürün ya da konuya ilişkin etkin bilgi toplamada ve karar verme sürecinde etkili bir biçimde rol alabilir.

Görüş Madenciliği alanında gerçekleştirilen çalışmalar, makine öğrenmesine dayalı yöntemler ve sözlüğe dayalı yöntemler olmak üzere iki temel sınıf altında incelenebilir [2]. Makine öğrenmesine dayalı yöntemler ile görüş madenciliği gerçekleştirilirken, görüş kutbu etiketlenmiş veri seti makine öğrenmesi algoritmalarından birine (karar ağaçları, destek vektör makineleri, yapay sinir ağları, k-en yakın komşu, Naïve Bayes vb.) uygulanarak sınıflandırma modeli oluşturulmakta; ardından bu model yeni örneklerin sınıflandırılmasında kullanılmaktadır. Görüş madenciliğine ilişkin veri seti, dilsel özellikler kullanılarak temsil edilmektedir. Sözlüğe dayalı yöntemler ise görüş sınıflandırma işlemini daha önceden oluşturulmuş görüş sözcükleri içeren bir görüş sözlüğüne dayalı olarak gerçekleştirmektedir. Sözlüğe dayalı yöntemler, sözlük tabanlı ya da derlem tabanlı yaklaşımlar olmak üzere iki temel sınıf altında incelenmektedir. Sözlüğe dayalı yaklaşımlar ile metnin görüş kutbu belirlenirken, metinde geçen sözcük ve tümcelerin anlamsal yönelimlerine dayalı bir hesaplama gerçekleştirilir [3]. Derlem tabanlı yaklaşımlarda, istatistiksel ya da semantik yöntemler kullanılarak görüş kutbu belirlenir.

Makine öğrenmesine dayalı yöntemler ise öğreticili (supervised), yarı-öğreticili (semi-supervised) ve öğreticisiz (unsupervised) yöntemler olmak üzere üç temel sınıf altında incelenmektedir. Öğreticili öğrenme, gözetimli öğrenme ya da denetimli öğrenme, öğreticisiz öğrenme, gözetimsiz öğrenme ya da denetimsiz öğrenme, yarı-öğreticili öğrenme ise yarı-gözetimli öğrenme ya da yarı-denetimli öğrenme biçiminde de kullanılabilmektedir. Görüş madenciliğine ilişkin yöntemlerin temel sınıflandırması Şekil 1'de sunulmaktadır.

Sözlüğe dayalı görüş madenciliği yöntemleri, yüksek ölçeklenebilirlikleri nedeniyle görüş sınıflandırmada sıklıkla uygulanmaktadır. Ancak, makine öğrenmesi, sınıflandırma problemi için birçok etkin algoritma içermektedir. Makine

öğrenmesi algoritmalarının basit yapısı ve farklı alanlardaki görüş sınıflandırma uygulamalarına kısmen kolay uyarlanabilir yapısı, makine öğrenmesini görüş madenciliği için önemli bir araştırma alanı durumuna getirmektedir [2]. Makalenin bundan sonraki bölümünde, görüş madenciliğinde makine öğrenmesine dayalı kullanılan yöntemler, öğreticili, yarı-öğreticili ve öğreticisiz makine öğrenme yöntemleri sınıflandırmasına dayalı olarak sunulmaktadır.



Şekil 1: Görüş madenciliği yöntemlerinin sınıflandırılması [2].

2 Öğreticili öğrenme yöntemleri

Öğreticili öğrenme yöntemleri, temel makine öğrenmesi yöntemlerinin, derlemdeki alana özgü etiketler ile eğitilmesi ile sınıflandırma modelini oluşturur. Bu yöntemlerin, düzgün bir şekilde çalışabilmesi için, eğitim için kullanılan veri setinin yeterince büyük olması gerekmektedir. Öğreticili öğrenme yöntemlerine dayalı en temel çalışma Pang ve diğ. [4] tarafından gerçekleştirilen çalışmadır. Bu çalışmada, destek vektör makineleri, Naïve Bayes ve maksimum entropi sınıflandırıcıları film değerlendirmeleri içeren "Movie Review" veri seti üzerinde uygulanmıştır. Veri setinin temsilinde, 1-gram, 1-gram ve 2-gram, 1-gram ve sözcük türü etiketleri, sıfatlar gibi sekiz farklı özellik temsili yapısı, terim sıklığı ya da varlığına dayalı olarak ele almıştır. En yüksek doğru sınıflandırma oranının (%82.9) destek vektör makineleri ile veri 1-gram yöntemi ile temsil edildiğinde alındığı görülmüştür.

Pang ve Lee [5] çalışmalarında, belgede yer alan tümcelerin öznel ya da nesnel olarak sınıflandırılması ve nesnel olanların inceleme dışı bırakılması, kalan tümceler üzerinde makine öğrenmesi uygulanmasına yönelik bir yaklaşım önermiştir. Bu yaklaşım ile elde edilen iyileştirme öznel tümcelerin, nesnel tümcelere kıyasla ayırt ediciliği ve bilgi vericiliği daha yüksek özellikler içerdiğini göstermektedir. Bu çalışmada, "Movie Review" veri seti için elde edilen doğru sınıflandırma yüzdesi %87'ye çıkmıştır.

Görüş metinlerinin temsilinde sıklıkla görüş içeren veri setinin kelime torbası kullanılarak ya da metinde geçen sözcüklerin görülme sıklıklarına dayalı olarak temsil edildiği görülmektedir. Bunun yanı sıra, bazı çalışmalarda metinde yer alan her bir sözcüğün "iyi" ya da "kötü" olarak gruplandırılmasına dayalı olarak metne ilişkin genel bir iyi/kötü puan hesaplanması ile görüş sınıflandırma yapılabilmektedir. Ancak, bu temsil yöntemlerinin hiçbirisi metinde yer alan anlamsal bağlantıları yansıtabilmek için tam olarak yeterli değildir [6]. Bu nedenle, Whitelaw ve diğ. [6] tarafından gerçekleştirilen çalışmada, belirli bir tutumu ifade etmede kullanılan kelime toplulukları aracılığı ile metinlerin anlamsal ilişkileri daha kapsamlı bir biçimde ele alınmıştır. Destek Vektör Makineleri, farklı değerlendirme grupları ve kelime torbası özellikleri ile eğitilerek, "Movie Review" veri seti üzerinde %90.2 oranında bir doğru sınıflandırma oranı elde edildiği gözlenmiştir.

Matsumoto ve diğ. [7] çalışmalarında, belge seviyesi görüş sınıflandırma için, tümcelerde geçen sözcük sırası ve sentaks ilişkilerini göz önünde bulunduran bir yöntem önermiştir. N-gram temsili kullanılarak yalnızca birbirini takip eden sözcüklerin birlikte görülme durumları modellenebilmektedir. Bu çalışmada ise, aynı tümcede geçen ancak birbirlerini takip etmeyen sözcüklerin de birlikte görülme durumları dikkate alınmıştır. Örneğin, "The film however is all good" tümcesi N-gram yöntemi ile temsil edildiğinde, "film" ve "good" sözcükleri arasındaki ilişki gösterilemezken; sözcük alt-dizinlerine dayalı yöntem kullanılarak bu sözcükler arasındaki ilişki ("film-good") temsil edilebilmiştir. Tümcelerin temsilinde, ağaç yapısı kullanılarak, tümcede geçen sözcükler arasındaki ilişkiler, bağlar ile belirtilmiştir. Bu nedenle, N-gram temsiline kıyasla, daha zengin bir temsil sağlanmıştır. Geliştirilen yöntem ile %93.7 doğru sınıflandırma oranı elde edilmiştir.

McDonald ve diğ. [8] çalışmalarında, metinde belirtilen görüşleri farklı detay seviyelerinde (tümce seviyesi ya da belge seviyesi) sınıflandırabilecek ve bir seviyede gerçekleştirilen sınıflandırmanın diğer seviyedeki kararları da etkilediği bir yapı sunmuştur. Bu modelde, Viterbi algoritması kullanılmıştır. Tümce ve belge seviyesi bilgiyi bir araya getiren bir model tasarlanması, hem yalnızca belge seviyesi hem de yalnızca tümce seviyesi görüş madenciliğine kıyasla daha iyi deneysel sonuçlar alınmasını sağlamıştır.

Yeterli etiketlenmiş verinin varlığında, öğreticili makine öğrenmesi yöntemleri, görüş madenciliği için etkin çözümler üretebilmektedir. Ancak, veri setine açıklamalar eklenmesi doğal dil işlemenin oldukça maliyetli etkinliklerinden biridir. Zaidan ve diğ. [9] çalışmalarında, bu gereksinimden hareketle, aktif ya da yarı-öğreticiliği öğrenme algoritmalarına alternatif bir yaklaşım olarak, insanların az sayıda eğitim verisine açıklamalar koymasını önermiştir. Bu öneride, insanların belirli bir ifadenin pozitif ya da negatif olduğunu belirlemede kullandıkları ifadeleri belirtmeleri istenmiş ve destek vektör makineleri ile görüş verilerinin sınıflandırılmasında, belgenin görüş kutbu etiketlerine ek olarak, insanlardan elde edilen ek bilgiler kullanılmıştır. Böylelikle, görüş kutbuna ilişkin karara ulaşmada kullanılan gerekçeler, eğitim seti olarak verilmiştir. Yöntem, destek vektör makineleri ile "Movie Review" veri setinde uygulandığında, %92.2 doğru sınıflandırma oranı elde edildiği görülmüştür.

Tan ve Zhang [10] çalışmalarında, Çince belgelerde görüş madenciliği için, dört farklı özellik seçimi ve beş öğrenme algoritmasının etkinliğini araştırmıştır. Özellik seçim yöntemi olarak, eğitim derlemindeki her bir sözcüğün sıklığını dikkate alan belge sıklığı ölçütü, ki-kare ölçütü, karşılıklı bilgi ve bilgi kazancı yöntemleri uygulanmıştır. Sınıflandırma yöntemi olarak ise, kitle merkezi sınıflandırıcısı, k-en yakın komşu sınıflandırıcısı, destek vektör makineleri, Naïve Bayes ve Winnow sınıflandırma yöntemleri kullanılmıştır. Deneysel çalışmalarda 458'i pozitif, 563'ü negatif toplam 1021 Çince belgeden oluşan bir derlem, geri çağırma, kesinlik ve F1 ölçütlerine dayalı olarak değerlendirilmiştir. Deneysel sonuçlara göre en iyi sonuçlar bilgi kazancı, özellik seçimi ve destek vektör makinesi sınıflandırıcısı ile elde edilmiştir. Bunun yanı sıra, bir alanda eğitilen bir sınıflandırıcının başka bir alana ait görüş madenciliği veri seti ile test edilmesinde, sınıflandırıcı performanslarının önemli ölçüde düştüğü

gözlenmiştir. Dolayısı ile deneysel sonuçlar görüş sınıflandırıcılarının önemli ölçüde alan ve konu bağımlı olduğunu göstermektedir.

Prabowo ve Thelwall [11] çalışmalarında, görüş sınıflandırma için kural tabanlı sınıflandırma, öğreticili öğrenme ve makine öğrenmesi tekniklerini bir arada kullanan melez bir sınıflandırma mimarisi sunmaktadır. Bu mimaride, sınıflandırıcılar, birbirleri ile ardışık olarak konuşlandırılarak, bir sınıflandırıcının belgeyi sınıflandırmaması durumunda, bir sonraki sınıflandırıcıdan devam edilmesi şeklinde bir yapı benimsenmiştir. Çalışmada, genel soruşturmaya dayalı sınıflandırma, kural tabanlı sınıflandırma, istatistik tabanlı sınıflandırma ve tümevarım kural tabanlı sınıflandırma yöntemleri kullanılmıştır. Genel soruşturmaya dayalı sınıflandırmada, "General Inquirer Lexicon"da yer alan 3672 ön-sınıflandırılmış (1598: pozitif, 2074: negatif) sözcük kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. Kural tabanlı sınıflandırıcı da belge seti için tümcede bulunan her bir nesne "?" ya da "#" ile değiştirilerek, kuralların koşul kısımları oluşturulmuştur. Ardından, her bir koşula uygun duygu belirteci atanarak kurallar oluşturulmakta ve kural tabanlı sınıflandırıcı uygulanarak sınıflandırılmaktadır. Belgede yer alan tümcelerin çözümlenmesinde "Montylingua Chunker" aracı kullanılmıştır. Çözümlemesi gerçekleştirilen tümcelerde "NNP" ya da "NNPS" olarak etiketlenen her bir nesne "?" ile değiştirilmiştir. İstatistik tabanlı sınıflandırıcı, her bir kuralın koşul kısmı ile duygu belirten sözcükler arasındaki yakınlığı, istatistiksel ölçütlere dayalı olarak ölçümlemiştir. Deneysel çalışmalar, farklı sınıflandırma yöntemleri art arda uygulanarak gerçekleştirilmiştir. Deneysel sonuçlarda, istatistik tabanlı sınıflandırıcının, melez sınıflandırıcının etkinliğini artırdığı, genel soruşturmaya dayalı sınıflandırıcının, melez sınıflandırıcının etkinliğini her zaman artırmadığı; bazen, daha kötü sonuçlar alınmasına neden olduğu, destek vektör makineleri sınıflandırıcının yeterli pozitif ve negatif örneklerin varlığında iyi sonuçlar verdiği görülmüştür.

Yessenalina ve diğ. [12] çalışmalarında, belge seviyesi görüş madenciliği için çok katmanlı bir öğreticili yaklaşım sunmuştur. Belirli bir metinde yer alan tümcelerin görüş sınıflandırmaya ilişkin ayırt ediciliği ya da bilgi vericiliği aynı değildir. Bu nedenle, bu yaklaşım ile görüş madenciliği bakımından bilgi vericiliği/ayırt ediciliği yüksek tümcelerin belirlenmesi ve bu tümceler üzerinde görüş madenciliği yapılması amaçlanmıştır. Bu yöntemde, tümce seviyesi etiketler saklı değişkenler olarak kullanılmakta, böylelikle tümce-seviyesi açıklama ekleme maliyeti ortadan kalkmaktadır. Önerilen yaklaşım, "Movie Review" veri seti üzerinde, destek vektör makineleri yöntemi ile uygulandığında, %93.22 doğru sınıflandırma yüzdesi elde edildiği gözlenmiştir.

Qui ve diğ. [13] çalışmalarında, çevrimiçi reklamcılık için görüş madenciliğine dayalı bir yöntem sunmuştur. Görüş madenciliği ile reklamların hem konu bakımından ilgili olması, hem de kullanıcı tutumlarını göz önünde bulunduracak şekilde konuşlandırılması amaçlanmıştır. Geliştirilen yöntemde öncelikle web sayfasından konu sözcükleri elde edilmektedir. Ardından, tüketicilerin ilgili konu sözcüklerine ilişkin görüşleri, görüş sınıflandırma aracılığı ile elde edilmekte ve reklamcılığa ilişkin anahtar kelimeler uygun reklamlar ile eşleştirilmektedir. Geliştirilen yöntemde, sözdizimsel ayrıştırma (syntactic parsing), görüş sözlüğü ve kural tabanlı sınıflandırma yöntemleri kullanılarak konu sözcükleri ve tüketici tutumları belirlenmiştir. Çalışma, web forumları üzerinden elde edilen veri seti üzerinde uygulanmıştır.

Zhao ve diğ. [14] çalışmalarında, görüş sınıflandırma için iki farklı öznitelik yöntemi önermiştir. Tümcelerin görüşlerinin belirlenmesinde, her zaman tümce içi özellikler yeterli olmamaktadır. Tümce seviyesinde görüş sınıflandırma performansının iyileştirilmesinde, tümce dışı bazı özniteliklerin de göz önünde bulundurulması yararlı olabilmektedir. Bu nedenle, bu çalışmada, belge içi ve belgeler arası öznitelikler kullanılmaktadır. Ardından, çizge-tabanlı algoritma kullanılarak, bu özniteliklerin yayılımı sağlanmaktadır. Geliştirilen yöntem, "Camera" veri seti üzerinde uygulanarak, öğreticisiz, destek vektör makineleri ve maksimum entropi sınıflandırıcıları ile karşılaştırılmış ve %67.23 doğru sınıflandırma oranı elde edildiği gözlenmiştir.

Bai [15] çalışmasında, görüş sınıflandırma için Markov model sınıflandırıcılarına dayalı iki aşamalı bir algoritma tasarlamıştır. Markov model sınıflandırıcı ile sözcükler arasındaki bağlantılar belirlenmiş, görüş sözlüğü de kullanılarak temel sınıflandırma yöntemlerinin performanslarında iyileştirme sağlanmıştır. Sözcükler arasındaki bağlantılar Markov modeli ile modellendikten sonra, Tabu arama algoritması kullanılarak algoritmanın parametreleri en iyileştirilerek daha yüksek bir çapraz geçerleme doğruluk oranı elde edilmektedir. Deneysel çalışmalarda, "IMDb v1.0", "IMDb v2.0", "M&A News", "Financial News" ve "Mixed News" veri setleri kullanılmıştır. Geliştirilen algoritma, destek vektör makineleri, Naïve Bayes ve maksimum entropi gibi başlıca sınıflandırma algoritmaları ile karşılaştırılmıştır. Algoritma ile film değerlendirmelerine ilişkin veri setinde %92.7 doğru sınıflandırma başarımı elde edilmiştir.

Tablo 1'de bu bölümde sunulan öğreticili öğrenme yöntemlerine ilişkin karşılaştırma sunulmaktadır. Burada, her bir çalışmada kullanılan temel yöntemler, geliştirilen yöntemlerin uygulandıkları veri alanları, değerlendirmede kullanılan temel ölçütler ve elde edilen en iyi sonuçlar özetlenmektedir.

Chen ve Tseng [16] çalışmalarında, görüş ifadelerinin niteliklerinin belirlenmesi işlemini sınıflandırma problemi olarak ele almıştır. Sınıflandırmada, iki farklı destek vektör makinesi yöntemi olan bire-karşı-tüm destek vektör makinesi (one-versus-all support vector machine-SVM) ve tek-makine çok-sınıflı destek vektör makinesi (single-machine multi-class support vector machine-SVM) yöntemleri kullanılmıştır. Sınıflandırmada kullanılacak veri setinin özniteliklerinin belirlenmesi için bilgi niteliği çatısı (information quality framework) kullanılmıştır. Geliştirilen mimari, "Digital Cameras" ve "MP3 Reviews" veri setleri üzerinde uygulanmıştır.

Wang ve diğ. [17] çalışmalarında, görüş sınıflandırma için fisher diskriminant oranına dayalı bir öznitelik seçim yöntemi önermiştir. Deneysel çalışmalarda, geliştirilen öznitelik seçim yöntemi, bilgi kazancına dayalı öznitelik yöntemi ile karşılaştırılmıştır. Sınıflandırma algoritması olarak, destek vektör makineleri yöntemi kullanılmıştır. Deneylerde Çince görüş ifadelerine ilişkin bir derlem olan "COAE2008" ve otomobillere ilişkin görüş ifadeleri içeren bir veri seti kullanılmıştır.

Tablo 1: Öğreticili öğrenme yöntemlerinin karşılaştırılması.

Referans	Yıl	Yöntemler	Veri Alanı	Değerlendirme Ölçütü	Değer
[4]	2002	Destek vektör makineleri, Naïve Bayes, maksimum entropi, N-gram temsili	Film Değerlendirmesi	Doğru Sınıflandırma	%82.90
[5]	2004	Tümcelerin öznel/nesnel olarak sınıflandırılması	Film Değerlendirmesi	Doğru Sınıflandırma	%87.00
[6]	2005	Tutum bildiren kelime toplulukları ile anlamsal ilişki temsili, Destek vektör makineleri	Film Değerlendirmesi	Doğru Sınıflandırma	%90.20
[7]	2005	Sözcük sırası ve sentaks ilişkilerine dayalı temsil, ağaç yapısı	Film Değerlendirmesi	Doğru Sınıflandırma	%93.70
[8]	2007	Viterbi algoritması, tümce/belge seviyesi sınıflandırma	Çevrimiçi Ürün Değerlendirmesi	Doğru Sınıflandırma	%82.80
[9]	2007	Görüş kutbu etiketleri, kişilerden elde edilen ek açıklamalar, Destek vektör makineleri	Film Değerlendirmesi	Doğru Sınıflandırma	%92.20
[10]	2008	Belge sıklığı ölçütü, ki-kare ölçütü, karşılıklı bilgi ve bilgi kazancı yöntemleri, kitle merkezi sınıflandırıcısı, k-en yakın komşu sınıflandırıcısı, destek vektör makineleri, Naïve Bayes, Winnow sınıflandırma yöntemleri	Çince Görüş Değerlendirmeleri	F1 Ölçütü (Makro/Mikro)	%86.64
[11]	2009	Kural tabanlı sınıflandırma, öğreticili öğrenme, makine öğrenmesi (Destek vektör makineleri)	Film ve Ürün Değerlendirmesi	F1 Ölçütü (Makro/Mikro)	%91.00
[12]	2010	Çok katmanlı öğreticili mimari, Tümce seviyesi etiketler, Destek vektör makineleri	Film Değerlendirmesi	Doğru Sınıflandırma	%93.22
[13]	2010	Sözdizimsel ayrıştırma, görüş sözlüğü, kural tabanlı sınıflandırma	Web forumları	Doğru Sınıflandırma	%55.00
[14]	2010	Belge içi ve belgeler arası öznitelikler, Çizge-tabanlı yayılım algoritması, destek vektör makineleri, maksimum entropi, öğreticisiz öğrenme	Kamera Değerlendirmesi	Doğru Sınıflandırma	%67.23
[15]	2011	Markov model, Görüş sözlüğü, Tabu arama algoritması, Destek vektör makineleri, Naive Bayes, maksimum entropi	Film Değerlendirmesi	Doğru Sınıflandırma	%92.70
[16]	2011	Bire-karşı-tüm destek vektör makinesi, tek-makine çok-sınıflı destek vektör makinesi, Bilgi niteliği çatısı	Ürün Değerlendirmeleri	F1 Ölçütü (Makro/Mikro)	%91.40
[17]	2011	Fisher diskirminat oranına, Bilgi kazancına dayalı öznitelik seçimi, Destek vektör makineleri	Çince Görüş Değerlendirmeleri	Doğru Sınıflandırma	%86.61
[18]	2011	Sınıflandırıcı toplulukları, Naïve Bayes, maksimum entropi ve destek vektör makineleri, sözcük tipi bilgisi, sözcük ilişkileri ve özellik ağırlıklandırma yöntemleri	Ürün Değerlendirmeleri	Doğru Sınıflandırma	%88.65
[19]	2012	Görüş Sözlüğü, iyileştirilmiş Naïve Bayes	Restoran Değerlendirmeleri	Pozitif/Negatif Sınıf Doğru Sınıflandırma yüzdesi arası fark	%03.60
[21]	2013	Dilsel özelliklere dayalı öznitelik çıkarımı, TF-IDF terim puanlama, destek vektör makineleri	Sosyal Medya	Doğru Sınıflandırma	%90.40
[22]	2013	Destek vektör makineleri, yapay sinir ağları, bilgi kazancı öznitelik çıkarımı	Ürün Değerlendirmeleri	Doğru Sınıflandırma	%90.30
[23]	2013	Naïve Bayes, maksimum entropi, karar ağacı, k-en yakın komşu, Destek vektör makineleri, bagging, boosting ve random subspace yöntemleri	Film ve Ürün Değerlendirmesi	Doğru Sınıflandırma	%92.20

Geliştirilen yöntem ile ilgili veri setleri ile sırasıyla %86.61 ve %82.80 doğru sınıflandırma oranları elde edildiği görülmüştür.

Xia ve diğ. [18] çalışmalarında, görüş sınıflandırma için sınıflandırıcı toplulukları kullanılmasının etkinliğini incelemektedir. Sınıflandırıcı toplulukları, farklı özellik setlerini ve sınıflandırma algoritmalarını bir araya getirerek yüksek sınıflandırma başarımı elde edilmesini amaçlamaktadır. Çalışmada sınıflandırma algoritması olarak Naïve Bayes, maksimum entropi ve destek vektör makineleri yöntemleri; özellik seçimi yöntemi olarak, sözcük tipi bilgisi, sözcük ilişkileri ve özellik ağırlıklandırma yöntemleri kullanılmıştır. Deneysel sonuçlar, "Movie", "Book", "DVD", "Electronic" ve "Kitchen" veri setleri üzerinde alınmıştır. Deneysel sonuçlar, özellik seçim setlerinin bir arada kullanılmasının tek bir sınıflandırma algoritması ile genellikle sınıflandırıcı topluluklarının kullanılmasına kıyasla daha iyi sonuçlar verdiğini göstermektedir.

Kang ve diğ. [19] çalışmalarında, restoranlara ilişkin değerlendirmeler üzerinde görüş sınıflandırma yapmak üzere, görüş sözlüğü ve iyileştirilmiş Naïve Bayes algoritmasına dayalı bir yöntem geliştirmiştir. Sınıflandırmada kullanılan veri setinin, dengesiz sınıf dağılımları göstermesi, sınıflandırma performansını olumsuz olarak etkileyen faktörlerin başında gelmektedir [20]. Kang ve diğ. [19] tarafından restoran değerlendirmeleri üzerinde gerçekleştirilen çalışmada, veri dengesizliği problemini ortadan kaldırmak amacıyla Naïve Bayes yöntemi iyileştirilmiştir. Standart Naïve Bayes yöntemi ile elde edilen pozitif sınıf doğru sınıflandırma yüzdesi ile negatif sınıf doğru sınıflandırma yüzdesi arası fark %28.5 iken, veri temsilinde 1-gram ve 2-gram kullanılması ile iyileştirilmiş Naïve Bayes yöntemi ile bu fark %3.6 oranına düşürülmüştür.

Li ve Li [21] çalışmalarında, mikroblog platformlarındaki görüşlere ilişkin sayısal bir özet sunan bir mimari sunmuştur. Geliştirilen mimarinin, görüş sınıflandırmaya ek olarak, güncel konu belirleme, güvenilirlik değerlendirme ve sayısal bir özet sunma işlevleri bulunmaktadır. Veri setinden özelliklerin çıkarımında eş anlamlı sözcüklere ve diğer bazı dilsel özelliklere ilişkin örüntü tanıma yöntemleri kullanılarak anlamsal örüntüler belirlenmiş; ardından, bu örüntüler TF-IDF terim puanlama yöntemi kullanılarak sıralanmıştır. Geliştirilen mimari, öncelikle mikroblog ortamlarından kullanıcı görüşlerine ilişkin verileri toplamakta; önişlem uygulanan veri seti veri tabanlarında tutulmaktadır. Güncel konu belirleme bileşeni, görüşler arasında değinilen, yorum belirtilen konuları belirlemektedir. Görüş sınıflandırma bileşeni ise destek vektör makineleri sınıflandırıcısını kullanarak görüşleri sınıflandırırken, negatif ve pozitif görüşlere ek olarak, görüş öznelliği ve güvenilirliği de dikkate alınmaktadır. Sistemin son çıktısı oluşturulurken, görüş öznelliği, görüş kutbu ile güvenilirlik değerini bir arada yansıtan bir değer oluşturulmaktadır. Geliştirilen yöntem, destek vektör makineleri sınıflandırıcısı ile "Twitter" veri seti üzerinde, 1-gram veri temsil yöntemi ve 5-kat çapraz geçerleme yöntemi aracılığıyla değerlendirildiğinde %90.4 doğru sınıflandırma oranı elde edilmiştir.

Moraes ve diğ. [22] çalışmalarında, belge-seviyesi görüş sınıflandırmada destek vektör makineleri ve yapay sinir ağları yöntemlerinin etkinliklerini deneysel olarak karşılaştırmıştır. Sınıflandırıcıların değerlendirilmesinde "Movie", "GPS", "Books" ve "Camera" veri setleri kullanılmış; doğru sınıflandırma, geri çağırma, kesinlik ve F-ölçütü hesaplatılmıştır. Deneysel çalışmada, 10-kat çapraz geçerleme yöntemi kullanılmıştır. Öncelikle, veri seti, Snowball kök ayırma yöntemi (Snowball stemmer) kullanılarak, köklerine ayrılmış; ardından, TF-IDF terim puanlama yöntemi kullanılarak sıralanmıştır. Özellik seçiminde, bilgi kazancı yöntemi kullanılmıştır. Deneysel sonuçlara göre, bilgi kazancı yöntemi kullanılarak binden fazla terimin özellik olarak seçilmesi, doğru sınıflandırma oranında ciddi bir artış sağlayamamaktadır. Test senaryolarının yarıya yakınında yapay sinir ağlarının destek vektör makineleri yönteminden daha iyi sonuçlar aldığı; ancak destek vektör makinelerinin yalnızca iki senaryo için yapay sinir ağlarını geçtiği görülmüştür. Dengesiz veri seti (imbalanced dataset) ile gerçekleştirilen deneyler, yapay sinir ağlarının, dengesiz ve gürültülü veriye karşı destek vektör makinelerine kıyasla daha duyarlı olduğunu göstermektedir.

Wang ve diğ. [23] çalışmalarında, görüş madenciliğinde sınıflandırıcı topluluklarının etkinliğini incelemek için, on farklı görüş madenciliği veri setini ("Camera", "Camp", "Doctor", "Drug", "Laptop", "Lawyer", "Movie", "Music", "Radio", "TV") altı farklı biçimde temsil ederek, beş farklı sınıflandırıcı ve üç farklı sınıflandırıcı topluluğu ile deneysel çalışmalar gerçekleştirmiştir. Temel sınıflandırıcı olarak, Naïve Bayes, maksimum entropi, karar ağacı, k-en yakın komşu ve destek vektör makineleri yöntemleri, sınıflandırıcı topluluğu yöntemleri olarak, Bagging, Boosting ve Random Subspace yöntemleri kullanılmıştır. Deneysel sonuçlar, görüş madenciliğinde, temel sınıflandırıcılar yerine, sınıflandırıcı topluluklarının kullanılmasının sınıflandırma başarımını önemli ölçüde iyileştirdiğini göstermektedir. Karşılaştırmada kullanılan yöntemler arasında en iyi sonuçlar Random Subspace sınıflandırıcı topluluğu yöntemi ile elde edilmiştir.

Chalothom ve Ellman [24] çalışmalarında, Twitter verileri üzerinde görüş sınıflandırma uygulaması gerçekleştirerek, destek vektör makineleri, Naive Bayes yöntemi ve sınıflandırıcı toplulukları yöntemlerinin etkinliklerini araştırmıştır. Bunun yanı sıra, veri setinin temsilinde sözcük torbası yöntemine ek olarak görüş sözlüğü kullanıldığında daha iyi F-ölçütü sonuçları elde edildiği gözlenmiştir.

Zheng ve diğ. [25] çalışmalarında, görüş madenciliğinde öznitelik seçiminin sınıflandırma algoritmalarının performansında etkinlikleri üzerine yoğunlaşmıştır. Bu doğrultuda, Çince çevrimiçi görüş değerlendirmelerine ilişkin veriler farklı N-gram temsil biçimleri ve sözcük türü birleşimleri ile incelenerek, en iyi sonuçların 4-gram temsili ile elde edildiği gözlenmiştir. Bunun yanı sıra, deneysel çalışmalar düşük N-gram temsillerinin, yüksek N-gram temsillerine kıyasla daha yüksek başarım elde ettiğini göstermektedir.

3 Yarı-öğreticili öğrenme yöntemleri

Yarı-öğreticili öğrenme yöntemleri, öğreticili öğrenme yöntemlerinin eğitim verisi ile test verisi arasındaki dağılım farklı olduğunda etkin sonuçlar vermemeleri, alana özgü olmaları; belirli bir alan için oluşturulan sınıflandırma modelinin, başka bir alanda uygulanabilmesi için sınıflandırıcının yeniden eğitilmesini gerektirmesi gibi kısıtlarını ortadan kaldırmak amacıyla görüş madenciliğinde sıklıkla uygulanmaktadır [26]. Yarı-öğreticili öğrenmede, büyük miktarda etiketsiz veri ile az sayıda etiketli veri, eğitim amacıyla kullanılmaktadır. Web'in alan ve içerik bakımından zengin yapısı, alana özgü etiketlemeyi ve derlem hazırlamayı zorlaştırmakta; görüş madenciliği için yarı-öğreticili yöntemler geliştirme gerekliliğini artırmaktadır. Görüş sınıflandırmada

yarı-öğreticili öğrenme yöntemleri, alan-transfer problemini ortadan kaldırmak için kullanılmaktadır. Alan-transfer problemi, bir alanda eğitilen bir sınıflandırıcının, eğitildiği alandan farklı bir alana ait veri setini sınıflandırmak için kullanıldığında performansının önemli ölçüde düşmesidir. Görüş madenciliği alanında üzerinde çalışılan problemlerden biri de farklı dillerdeki görüş metinlerinin sınıflandırılmasıdır. Bu problem, belirli bir dildeki etiketli veri seti kullanılarak eğitilen sınıflandırma algoritmasının, başka bir dildeki veri setinin sınıflandırılması amacıyla kullanılması durumunda görülmektedir. Görüş sınıflandırmada yarı-öğreticili öğrenme yöntemleri, alan-transfer problemine ek olarak, farklı dillerdeki görüş ifade eden metinlerin sınıflandırılması için de sıklıkla kullanılan yöntemler arasındadır.

Görüş sınıflandırmada yarı-öğreticili öğrenme yöntemlerine dayalı en temel çalışmalardan biri Aue ve Gamon [27] tarafından gerçekleştirilen çalışmadır. Bu çalışmada, yeterli etiketli verinin bulunmadığı alanlarda, sınıflandırıcı eğitimi için farklı stratejilerin etkinlikleri araştırılmıştır. Bu stratejilerden birincisi, sınıflandırıcının eğitilmesinde, etiketli verinin elde edilebildiği başka bazı alanlardan verilerin kullanılmasıdır. İkinci stratejide, birinci stratejideki gibi, eğitimde mevcut etiketli verilerden tümü kullanılmakta ancak özellik seti olarak yalnızca hedef alanda gözlenen özellikler dikkate alınmaktadır. Üçüncü stratejide, yeterli etiketli verinin elde edilebildiği alanlarda eğitilen sınıflandırıcılar, sınıflandırıcı toplulukları aracılığıyla bir araya getirilmektedir. Dördüncü stratejide ise az sayıda etiketli veri, çok sayıda etiketsiz veri ile bir arada kullanılarak eğitim gerçekleştirilmektedir. Deneysel çalışmalarda, dört farklı alandan, "Movie", "Book", "Product Support Services" ve "Knowledge Base" veri setleri kullanılmıştır. Veri setleri, 1-gram, 2-gram ve 3-gram yöntemleri aracılığıyla temsil edilmiştir. Deneylerde kullanılan dört farklı stratejinin tümünde, beklenti-maksimizasyonu (expectation-maximization) algoritmasının en yüksek doğru sınıflandırma oranını elde ettiği gözlenmektedir.

Tan ve diğ. [28] çalışmalarında, alan-transfer problemi (domain-transfer problem) için bir yöntem geliştirmiştir. Bu çalışmada, bir alanda eğitilen sınıflandırıcı, yeni karşılaşılan alandaki en bilgilendirici n adet etiketsiz veriyi etiketlemek için kullanılmaktadır. Bu yöntemde, kosinüs benzerlik ölçütüne dayalı olarak pozitif ve negatif benzerlikler belirlenmektedir. Bir örneğin yüksek bir negatif benzerliğe sahip olması negatif sınıfta; yüksek bir pozitif benzerliğe sahip olması ise pozitif sınıfta yer aldığını göstermektedir. Bu gözleme dayalı olarak, iki farklı sıralama yöntemi geliştirilmiştir: benzerlik sıralama yöntemi ve bağıl benzerlik sıralama yöntemi. Benzerlik sıralama yönteminde, öncelikle tüm negatif benzerliğe sahip örnekler sıralanmakta ve bu örneklerden ilk $n/2$ tanesi negatif örnekler olarak alınmaktadır. Aynı işlem, pozitif örnekler için de yapılarak n adet örnekten oluşan veri seti oluşturulmaktadır. Bağıl benzerlik sıralama yönteminde ise orijinal benzerlik değerleri standartlaştırılarak kullanılmaktadır. Deneysel çalışmalarda, üç farklı alana ait "Computer Reviews", "Educational Reviews" ve "House Reviews" veri setleri kullanılarak, bir alanda eğitilen bir sınıflandırıcının, önerilen yöntem doğrultusunda bir başka alanda uygulanmasının etkinliği araştırılmıştır. Karşılaştırmalı çalışmalarda, TSVM (Transductive Support Vector Machine) yöntemi kullanılmıştır. Deneysel sonuçlar, önerilen yöntemler ile genellikle TSVM ile kıyasla daha başarılı sonuçlar elde edildiğini göstermektedir.

Blitzer ve diğ. [29] çalışmalarında, alan-transfer problemi için yapısal yazışma öğrenme (structural correspondence learning) yöntemine dayalı, kaynak alandaki etiketli verileri ve hedef ve kaynak alanlardaki etiketsiz verileri kullanan bir yöntem geliştirmiştir. Kaynak ve hedef alanlar arasındaki bağlantı temel özelliklere dayalı olarak gerçekleştirilmektedir. Bu nedenle, bu temel özelliklerin seçimi, öğrenme yönteminin etkinliği için önemli bir faktördür. Temel özelliklerin seçimi iki aşamalı olarak gerçekleştirilmektedir. Öncelikle hem kaynak hem de hedef alanda sık görülen sözcükler, aday temel özellikler olarak belirlenmektedir. Ardından, aday temel özellikler ve kaynak etiketleri arasındaki karşılıklı bilgi (mutual information) hesaplanarak, temel özellikler belirlenmektedir. Geliştirilen yöntem ile alanlar arası adaptasyon bağıl hata oranı, öğreticili temel sınıflandırma algoritmasına kıyasla %46 oranında azaltılmıştır.

Mihalcea ve diğ. [30] çalışmalarında, farklı dillerdeki görüş metinlerinin sınıflandırılması için, İngilizce için geliştirilen mevcut görüş sözlüğü ve derlemlerin kullanılabilirliğini araştırmıştır. Çalışmada, mevcut İngilizce görüş sözlüğü ve sözlük kullanılarak, yeni bir dil için görüş sözlüğü oluşturulup oluşturulamayacağı incelenmiştir. İngiliz dili için geliştirilmiş duygu analizi araçları ve paralel derlem kullanılarak, yeni bir dil için görüş açıklamaları içeren bir derlem türetilip türetilemeyeceği ve yeni bir dilde görüş sınıflandırma yapabilmek için bu otomatik kaynakların kullanılabilirliği araştırılmıştır. Çalışmada, Romancada görüş sınıflandırabilmek için sözlük tabanlı ve derlem tabanlı yöntemler önerilmiştir. Sözlük tabanlı yöntemde İngilizce görüş sözlüğü, Romancaya çevrilerek kural tabanlı sınıflandırıcı kullanılarak sınıflandırılmıştır. Görüş madenciliğinde kullanılan ipuçlarının Romancada, orijinal sete kıyasla daha az güvenilir olduğu görülmüştür. Derlem tabanlı yöntemde ise İngilizce öznellik belirleme sınıflandırıcısı ve manuel olarak çevrimi gerçekleştirilmiş paralel derlem kullanılarak görüş sınıflandırma işlemi gerçekleştirilmiştir. Deneysel sonuçlar, derlem tabanlı yöntemin, sözlük tabanlı yönteme kıyasla daha iyi sonuçlar verdiğini göstermektedir.

Li and Zong [31] çalışmalarında, çok-alanlı görüş sınıflandırma (multi-domain sentiment classification) için iki farklı yöntem önermiştir. Bu yöntemlerden birincisinde, farklı alanlardaki özellik setleri bir araya getirilerek, sınıflandırma algoritmasının eğitilmesinde kullanılmak üzere tek bir özellik seti oluşturulmaktadır. İkinci yöntemde ise, farklı alanlardaki veri setleri üzerinde eğitilen temel sınıflandırıcılar meta öğrenme algoritması kullanılarak birleştirilmektedir. Deneysel çalışmalar, dört farklı alana ait "Books", "DVDs", "Electronics" ve "Kitchen Appliances" veri setleri üzerinde gerçekleştirilmiştir. Deneysel çalışmalar, tek bir temel sınıflandırıcıya kıyasla geliştirilen yöntemler ile daha yüksek sınıflandırma başarımı elde edildiğini göstermektedir. Bunun yanı sıra, sınıflandırıcıları bir araya getiren yöntem (ikinci yöntem) genellikle özellikleri bir araya getiren yöntemden (birinci yöntem) daha iyi sonuçlar vermektedir.

Banea ve diğ. [32] çalışmalarında, İngiliz dili için geliştirilmiş kaynakları, farklı iki dilde (Romanca ve İspanyolca) görüş öznellik sınıflandırması yapabilmek için kullanabilmek amacıyla makine çevirisine (machine translation) dayalı bir yöntem önermiştir. Deneysel çalışmalarda üç farklı senaryo gerçekleştirilmiştir. Birinci senaryoda, kaynak dilde manuel olarak görüş bildiren açıklamalar eklenmiş derlem makine çevirisi tekniklerine dayalı olarak hedef dile çevrilmiştir. Burada kaynak dilde görüş bildiren etiketler çıkarılarak, bu etiketler hedef dile çevrilmiştir. İkinci senaryoda, kaynak dildeki metin, kaynak dildeki görüş bildiren açıklama ekleme

aracı kullanılarak etiketlenmiş, ardından hedef dile çevirme işlemi gerçekleştirilmiştir. Üçüncü senaryoda ise hedef dildeki metin, makine çevirisi teknikleri aracılığıyla kaynak dile çevrilerek, kaynak dildeki görüş bildiren açıklama ekleme aracı kullanılarak etiketlenmekte ve sınıflandırma işlemine tabi tutulmaktadır. Deneysel çalışmalar, Naïve Bayes ve destek vektör makineleri sınıflandırıcıları kullanılarak gerçekleştirilmiş, birinci senaryonun, ikinci ve üçüncü senaryoya kıyasla daha düşük başarım gösterdiği; en yüksek başarım oranının ikinci senaryo için %69.44 F-ölçütü değeri ile ve üçüncü senaryo için %67.86 F-ölçütü değeri ile elde edildiği görülmektedir. Deneysel sonuçlar, makine çevirisinin, hedef dilde görüş bildiren açıklama ekleme amacıyla kullanılabilecek etkin yöntemlerden biri olduğunu göstermektedir.

Li ve diğ. [33] çalışmalarında, yarı-öğreticili görüş sınıflandırma için negatif olmayan matris faktorizasyonuna dayalı bir yöntem geliştirmiştir. Burada öğrenme, alan bağımsız görüş sözlüğünün, alan bağımlı etiketsiz veri ve az sayıda etiketli veri ile birlikte kullanımı ile gerçekleştirilmektedir. Deneysel sonuçlar, dört farklı alana ait veri setleri ile gerçekleştirilmiştir ve öğreticili sınıflandırma yöntemleri ile karşılaştırılabilir sonuçlar alındığını göstermektedir. Deneysel sonuçlar, görüş sınıflandırmada alan bağımsız ve alana özgü bilgilerin dikkate alınmasının, sınıflandırma başarımını artırdığını göstermektedir.

Dasgupta ve Ng [34] çalışmalarında, spektral kümeleme (spectral clustering), aktif öğrenme (active learning) ve transdüktif öğrenme (transductive learning) yöntemlerine dayalı bir yarı-öğreticili yöntem geliştirmiştir. Öncelikle, spektral kümeleme algoritması kullanılarak, temel (belirli) görüş ifadeleri belirlenmektedir. Ardından, aktif öğrenme algoritması kullanılarak, belirsiz görüş ifadeleri etiketlenmektedir. Daha sonra, transdüktif öğrenme algoritması kullanılarak geriye kalan görüş ifadeleri etiketlenmektedir. Deneysel çalışmalarda, beş farklı veri seti kullanılmaktadır. Geliştirilen yöntem, yarı-öğreticili spektral kümeleme, transdüktif destek vektör makineleri ve aktif öğrenme ile karşılaştırılmıştır. Deneysel sonuçlar, kümeleme algoritması kullanılarak, etiketlenmesi gereken görüş ifadesi sayısının azaltılmasının, yüksek başarımlı görüş sınıflandırma sistemleri geliştirmek için etkin bir yöntem olduğunu göstermektedir.

Wan ve diğ. [35] çalışmalarında, görüş sınıflandırma işlemini Çince değerlendirmeler üzerinde yapabilmek için İngilizcedeki gibi yeterince kaynak (derlem, sözlük vs.) olmaması nedeniyle, eş-eğitim (co-training) öğrenme algoritmasına dayalı bir yöntem geliştirmiştir. Birinci yöntem, etiketli İngilizce görüş ifadeleri kullanılarak sınıflandırma algoritmasının eğitilmesidir. Ardından, Çince görüş bildiren metinler makine çevirisi kullanılarak İngilizceye çevrilmekte ve etiketli İngilizce görüş ifadeleri kullanılarak eğitilen sınıflandırıcı aracılığıyla sınıflandırılmaktadır. İkinci yöntemde ise öncelikle etiketli İngilizce değerlendirme ifadeleri Çinceye çevrilmekte ardından sınıflandırıcının eğitimi, Çinceye çevrilmiş veri üzerinde gerçekleştirilerek, etiketsiz görüş ifadelerinin sınıflandırılması için kullanılmaktadır. Ancak, değinilen iki yöntem tek başlarına uygulandıklarında, Çince için yeterince yüksek doğru sınıflandırma oranı elde edilememektedir. Dolayısı ile bu çalışmada, eş-eğitim algoritması uygulanarak, İngilizce ve Çince özellikle bütünleşik bir yapıda bir arada kullanılarak daha yüksek bir doğru sınıflandırma oranı elde edilmektedir.

Li ve diğ. [36] çalışmalarında, görüş sınıflandırma için kişisel görüşe dayalı ve kişisel görüşe dayalı olmayan iki farklı model uygulamıştır. Kişisel görüşe dayalı modelde, herhangi bir kişinin belirli bir hedef nesneye ilişkin duygu, tercih, tutum vb. bildiren ifadeleri yer alırken; kişisel görüşe dayalı olmayan modelde, nesnel alana-özgü bilgi yer almaktadır. Deneysel sonuçlar, öğreticisiz öğrenme ile görüş sınıflandırma için iki boyutlu (kişisel görüşe dayalı ve kişisel görüşe dayalı olmayan) bir modelin uygulanması ile eş-öğretim (co-training) yarı-öğreticili öğrenme algoritması ile önemli performans iyileştirmeleri elde edildiğini göstermektedir.

He ve Zhou [37] çalışmalarında, makine öğrenmesine dayalı ve sözlüğe dayalı yöntemlerin sırasıyla yüksek veri etiketleme maliyetleri ve alan bağımlılıkları problemlerini ortadan kaldırmak amacıyla, kendi kendine eğitim (self-training) yaklaşımına dayalı bir yöntem sunmuşlardır. Bu yöntemde, sınıflandırma yöntemi öncelikle veri sözlüğünde önceden görüş kutupları atanmış sözcüklere dayalı olarak eğitilmektedir. Ardından, kendi kendine eğitim yaklaşımına dayalı özellik çıkarma yöntemi kullanılarak, alana özgü özellikler otomatik olarak elde edilmektedir. Bir başka sınıflandırma yöntemi, alana özgü elde edilen özellikler kullanılarak eğitilmekte ve modelin test edilmesinde kullanılmaktadır. Geliştirilen yöntem, "Movie Review" (MR) ve "Multi-domain sentiment" (MDS) veri setleri kullanılarak değerlendirilmiş ve mevcut yarı-denetimli öğrenme yöntemleri ile karşılaştırılabilir ya da daha iyi sonuçlar alındığı gözlenmiştir. Geliştirilen yöntem ile veri setlerinden, %70-75 aralığında doğru sınıflandırma yüzdeleri elde edilmiştir.

He ve diğ. [38] çalışmalarında, görüş kutbu belirten konuları, alan bağımsız görüş sözlüğü kullanarak ortak görüş-konu modeline (joint sentiment-topic model) dayalı olarak belirlemektedir. Görüş-konu modeli, gizli Dirichlet tahsis modeline (latent Dirichlet allocation model) dayanmaktadır ve metinden eş zamanlı görüş ve konu belirlemede kullanılmaktadır. Görüş-konu modeli, sınıflandırma işlemini gerçekleştirebilmek için yalnızca alan bağımsız görüş sözlüğü kullanılmasını gerektirmektedir. Alan bağımsız görüş sözlüğü kullanımı ile görüş-konu modeli ile "Movie Review" veri setinde %74, "Multi-domain sentiment" veri setinde %71 oranında doğru sınıflandırma başarımı elde edilmektedir. Görüş-konu modelinin eğitim için etiketli veriye gereksinim duymaması, yöntemi görüş sınıflandırmada alan-transfer probleminin ortadan kaldırılması için uygun bir yöntem durumuna getirmektedir. Bu çalışmada, öncelikle görüş kutbu belirten konular görüş-konu modeli aracılığıyla belirlendikten sonra, orijinal belgelere görüş kutbu belirtilen konular eklenerek yeni bir belge temsili elde edilmektedir. Ardından, bu yeni veri temsili maksimum entropi sınıflandırıcısı kullanılarak sınıflandırılmaktadır. Geliştirilen yöntem, "Movie Review" ve "Multi-domain sentiment" veri setlerindeki doğru sınıflandırma oranını %90'lara çıkarmıştır.

Tablo 2'de bu bölümde sunulan yarı-öğreticili öğrenme yöntemlerine ilişkin bilgiler özetlenmektedir.

Hernandez ve diğ. [39] çalışmalarında, bir metni görüş kutbuna dayalı olarak yalnızca pozitif ya da negatif olarak iki sınıfa ayırmak yerine, birbiriyle ilişkili üç farklı hedef değişkeninin (öznellik, görüş kutbu ve etkileme durumu) belirlenmesini amaçlamıştır. Bu nedenle, problem çok boyutlu bir sınıflandırma problemi şeklinde ele alınmış ve görüş madenciliği için Bayes ağları sınıflandırıcısına dayalı çok

boyutlu bir sınıflandırma modeli sunmuştur. Bunun yanı sıra, geliştirilen sınıflandırma modelini görüş madenciliği alanındaki etiketsiz veri setleri ile uygulayabilmek için beklenti-maksimizasyonu (expectation-maximization) algoritmasına dayalı bir yarı-denetimli öğrenme yöntemi uygulanmıştır.

Görüş sınıflandırma alanındaki kaynakların önemli bir bölümünün İngiliz dili için geliştirilmiş olması, farklı dillerde görüş madenciliği yapmak için otomatik makine çevirisine dayalı yöntemlerin kullanılmasını yaygınlaştırmıştır. Ancak, çeviri aracılığı ile elde edilen metin ile orijinal metin arasındaki terim dağılımların farklılığı ve çeviri kaynaklı hatalar, bu yöntemlerin doğru sınıflandırma başarım oranlarının yeterince yüksek olmasını engellemektedir. Hajmohammadi [40] çalışmalarında, görüş sınıflandırmada değinilen problemlerin ortadan kaldırılabilmesi için aktif öğrenme (active learning) ve yarı-öğreticili eş-eğitim algoritmalarına dayalı bir öğrenme modeli geliştirmiştir. Bu modelde, etiketleme maliyetini en aza indirgemek için, kaynak dildeki etiketli belgeler makine çevirisi yöntemi kullanılarak hedef dilde etiketli belgelere ve hedef dildeki etiketsiz belgeler makine çevirisi yöntemi kullanılarak kaynak dildeki etiketsiz belgeler dönüştürülerek iki farklı veri modeli oluşturulmaktadır. Yöntemin öğrenme aşamasında, iki sınıflandırıcı, daha önce değinilen şekilde oluşturulan veri setleri ile eğitilmekte ve ardından karşılık gelen modeldeki etiketsiz verilerin sınıflandırılması için kullanılmaktadır. Ardından, uyuşmazlık ve yoğunluk analizine dayalı olarak, insan etiketlemesi ile sınıflandırılan örneklerin karşılaştırması yapılmakta ve bu karşılaştırma sonucunda daha bilgilendirici olan örnekler seçilmektedir. Eş zamanlı olarak, eş-öğrenme algoritması doğru sınıflandırılmış örnekler arasından bazılarını seçmektedir. Bu iki farklı yöntem ile elde edilen örnekler bir sonraki döngüde eğitim verisine eklenerek, eğitim verisi iyileştirilmektedir. Önerilen yöntem, "Book Review" veri setinin üç farklı dildeki çevirileri ile değerlendirilmiştir.

Hajmohammadi [41] çalışmalarında, farklı dillerde görüş madenciliği için, Hajmohammadi [40]'da sunulan bir hedef ve bir kaynak dil içeren veri modelini iki kaynak dil ve bir hedef dil kullanılacak şekilde genişletmiş ve yarı-öğreticili öğrenme algoritması ile bütünleştirerek bir öğrenme modeli sunmuştur. Modelin eğitim aşamasında, her bir sınıflandırıcı, farklı bir veri modelindeki eğitim belgeleri ve etiketsiz belgeler kullanılarak eğitime tabi tutulmakta, ardından bu sınıflandırıcıların çıktıları, çoğunluğun azınlığı eğitmesi (majority teaching minority) kuralına dayalı olarak birleştirilmektedir.

Modelin test aşamasında, her bir sınıflandırıcı farklı bir veri modelindeki test belgeleri kullanılarak sınanmakta ve sınıflandırıcıların çıktıları, çoğunluk oylama (majority voting) yöntemi kullanılarak birleştirilmektedir. Deneysel çalışmalar, geliştirilen yöntemin, farklı dillerde görüş madenciliği uygulamak için uygun olduğunu göstermektedir.

Tablo 2: Yarı-Öğreticili öğrenme yöntemlerinin karşılaştırılması.

Referans	Yıl	Yöntemler	Veri Alanı	Değerlendirme Ölçütü	Değer
[27]	2005	Başka alandaki etiketli verilerin eğitimde kullanılması, Sınıflandırıcı Toplulukları, Beklenti-Maksimizasyonu	Film ve Ürün Değerlendirmeleri	Doğru Sınıflandırma	%82.39
[28]	2007	Kosinüs benzerlik ölçütü, Benzerlik Sıralama Yöntemi, Bağıl Benzerlik Sıralama Yöntemi, Transduktif Destek Vektör Makineleri	Bilgisayar, Eğitim ve Ev Değerlendirmeleri	Doğru Sınıflandırma	%89.93
[29]	2007	Yapısal yazışma öğrenme, karşılıklı bilgi ölçütü	Ürün Değerlendirmeleri	Doğru Sınıflandırma	%85.90
[30]	2007	İngilizce görüş sözlüğü, duygu analizi araçları, paralel derlem	Romanca Görüş Değerlendirmesi	F-ölçütü	%72.68
[31]	2008	Farklı alanlardaki özellik setlerinin bir araya getirilmesi, farklı alanlardaki veri setleri üzerinde sınıflandırıcı eğitimi, Meta Öğrenme	Ürün Değerlendirmeleri	Doğru Sınıflandırma	%85.00
[32]	2008	Makine Çevirisi, Açıklama Eklenmiş derlem, Görüş Açıklamaları Ekleme, Naïve Bayes, Destek vektör makineleri	Romanca Görüş Değerlendirmesi	F-ölçütü	%69.44
[34]	2009	Spektral Kümeleme, Aktif Öğrenme, Transdüktif Öğrenme	Film ve Ürün Değerlendirmeleri	Doğru Sınıflandırma	%76.20
[35]	2009	Eş-eğitim, İngilizce görüş ifadeleri, etiketli İngilizce değerlendirme ifadeleri	Çince Görüş Değerlendirmeleri	Doğru Sınıflandırma	%81.30
[36]	2010	Hedef nesneye ilişkin kişisel duygu, tercih, tutum ifadeleri, nesnel-alana özgü bilgiler, eş-öğretim	Ürün Değerlendirmeleri	Doğru Sınıflandırma	%86.75
[37]	2011	Kendi-kendine eğitim, Veri Sözlüğü, alana özgü özellikler	Film Değerlendirmesi, Çok Alanlı Görüş Değerlendirmesi	Doğru Sınıflandırma	%75.00
[38]	2011	Ortak görüş-konu modeli, gizli Dirchlet tahsis modeli, Eş Zamanlı Görüş ve Konu Belirleme	Film Değerlendirmesi, Çok Alanlı Görüş Değerlendirmesi	Doğru Sınıflandırma	%90.00
[39]	2013	Öznellik, Görüş Kutbu ve Etkileme Durumu Belirleme, Bayes Ağları, Beklenti Maksimizasyonu	Asomo veri seti	Doğru Sınıflandırma	%83.63
[40]	2014	Aktif Öğrenme, Yarı-öğreticili eş-eğitim	Kitap Değerlendirmesi	Doğru Sınıflandırma	%82.17
[41]	2014	Yarı-öğreticili öğrenme, çoğunluğun azınlığı eğitmesi kuralı	Kitap Değerlendirmesi	Doğru Sınıflandırma	%81.55

Hajmohammadi ve diğ. [42] çalışmalarında, İngiliz dili görüş kaynaklarının, diğer dillerdeki görüşlerin sınıflandırılmasında da etkin bir biçimde kullanılabilmesi için yarı-öğreticili öğrenme ve kendi kendine eğitim ilkelerine dayalı bir yöntem geliştirmiştir. Bu modelde, aktif öğrenme modelinin oluşturulmasında etkin örneklerin seçilebilmesi için etiketsiz verilerin yoğunluk analizi göz önünde bulundurulmuştur. Deneysel çalışmalarda İngiliz dilindeki kitap değerlendirmelerine ilişkin verilerin Çince, Japonca ve Fransızca dillerindeki görüş sınıflandırmada uygulanabilirliği araştırılmıştır. Deneysel sonuçlar, bir dildeki görüş kaynakları kullanılarak başka bir dile ait görüş ifadelerinin sınıflandırılmasında hedef dile ait etiketsiz verilerin de kullanılmasının, sınıflandırma başarımını önemli ölçüde artırdığını, aktif öğrenmenin otomatik etiketleme ile bir arada kullanılması ile öğrenme sürecinin önemli ölçüde hızlandırıldığını göstermektedir.

4 Öğreticisiz/Zayıf öğreticili yöntemler

Öğreticisiz/Zayıf Öğreticili öğrenme yöntemleri, öğreticili ya da yarı-öğreticili görüş madenciliği yöntemlerinde karşılaşılan alan bağımlılığı problemini ortadan kaldırmak için kullanılmaktadır. Öğreticisiz öğrenme yöntemlerinde, duygu sözlüğü, görüş kutbu belirlemek amacıyla ön bilgi olarak verilmektedir. Alan bağımsız duygu sözlükleri oluşturmanın alana özgü derlem oluşturmadan daha az maliyetli olmasından faydalanılmaktadır.

Öğreticisiz öğrenme alanındaki en temel çalışmalardan biri Turney ve Littman [43] tarafından gerçekleştirilen çalışmadır. Bu çalışmada, belirli bir belgenin pozitif ya da negatif olarak sınıflandırılması, belgede geçen belirteç ve sıfatların duygu yönlerine dayalı olarak gerçekleştirilmektedir. Duygu yönlerinin belirlenmesinde, yedisi pozitif (“good”, “nice”, “excellent”, “positive”, “fortunate”, “correct”, “superior) ve yedisi negatif (“bad”, “nasty”, “poor”, “unfortunate”, “wrong”, “inferior”) toplam on dört sözcük dikkate alınmıştır. Metinde geçen herhangi bir sözcüğün duygu yönü, değinilen yedi pozitif sözcüğe olan ilişkinin, yedi negatif sözcüğe olan ilişkiden çıkarılması ile hesaplanmaktadır. Sözcüklerin pozitif ve negatif sözcükler ile ilişkilerinin belirlenmesi için noktasal karşılıklı bilgi (pointwise mutual information) ve gizli anlamsal çözümleme (latent semantic analysis) yöntemleri kullanılmıştır. Noktasal karşılıklı bilgi ve gizli anlamsal çözümlemeye dayalı duygu yönü belirlemede farklı derlemler kullanılmıştır. Deneysel sonuçlar, gizli anlamsal çözümlemeye dayalı duygu yönü belirlemenin, karşılıklı bilgiye dayalı duygu yönü belirlemeye kıyasla veriyi daha etkin bir biçimde kullandığını ve yeterli büyüklüğe sahip bir derlem ile test edildiğinde, karşılıklı bilgiye dayalı duygu yönü belirleme ile elde edilen %80 başarım oranının aşılabileceğini göstermektedir.

Andreevskaia ve Bergler [44] çalışmalarında, farklı alanlarda tümce seviyesi görüş açıklamaları eklemek için, İngiliz dilinde geliştirilmiş bir sözcük veri tabanı olan WordNet üzerinde eğitilmiş sözlük-tabanlı bir sistem ile az miktarda etiketli, alana özgü veri ile eğitilmiş, derlem-tabanlı sınıflandırıcıyı bir araya getiren bir model geliştirmiştir. Deneysel çalışmalar, görüş sınıflandırma için tek başına sözlük bilgisi ya da eğitim verisinin kullanılmasına kıyasla, sözlük bilgisi ve derlem tabanlı sınıflandırıcı eğitimi kullanımı ile daha iyi sonuçlar alındığını göstermektedir.

Zagibalov ve Carroll [45] çalışmalarında, Çince ürün değerlendirmelerine ilişkin görüş ve nesnellik sınıflandırma için öğreticisiz bir yöntem sunmuştur. Bu yöntemde duygu yönlerinin belirlenmesi, Çincede “iyi” anlamına gelen tek bir sözcük içeren sözlük ile başlamaktadır. Bu sözlükte yer alan sözcük sayısı, kademeli olarak yeniden eğitim aracılığıyla artırılmaktadır. Eğitim sürecinde, yeni sözcüklerin sözlüğe eklenmesinde, bu sözcüklerin mevcut eğitim verisinin negatif ve pozitif kısımlarındaki kısmi sıklık değerleri gözetilmektedir. Herhangi bir metnin duygu yönü, duygu belirten sözcükler aracılığıyla hesaplanan duygu puanı ile belirlenmektedir. Bu yöntemde, yeniden eğitim sürecinin kaç kez tekrarlanacağı ve başlangıçta sözlükte hangi kelimenin yer alacağının belirlenmesi önemlidir. Aksi takdirde, yöntemin performansı, sözlüğe uygun olmayan sözcüklerin eklenmesi ile önemli ölçüde düşebilmektedir. Bu yöntem ile görüş kutbu belirleme için %87 F-ölçütü değeri elde edilmiştir.

Zagibalov ve Carroll [46] çalışmalarında, Çince ürün değerlendirmelerine ilişkin görüş sınıflandırma için öğreticisiz ve otomatik sözcük seçimine dayalı bir yöntem önermiştir. Bu yöntemde, otomatik sözcük seçimi sezgisel bilgiye dayalı olarak gerçekleştirilmektedir. Burada da [42]’de olduğu gibi sözlükte yer alan sözcük sayısı, kademeli olarak yeniden eğitim aracılığı ile artırılmaktadır. Tekrarlamalı yeniden eğitim süreci, herhangi bir belgenin sınıflandırması iki yineleme boyunca aynı kaldığı takdirde sonlandırılmaktadır. Geliştirilen yöntem ile öğreticili öğrenme yöntemlerine dayalı görüş sınıflandırmaya yakın ya da daha yüksek sonuçlar elde edilmiştir. Bu yöntem ile görüş kutbu belirleme için elde edilen en yüksek F-ölçütü %92’dir.

Qiu ve diğ. [47] çalışmalarında, görüş sınıflandırma için derlem tabanlı ve sözlük tabanlı yaklaşımları bütünleştiren bir öğreticisiz öğrenme modeli sunmuştur. Bu modelde de başlangıçtaki görüş sözlüğünün genişletilmesi tekrarlamalı bir sürece dayalı olarak gerçekleştirilmiştir. Ancak, burada görüş sözlüğü oluşturulmasına daha önceki yaklaşımlara kıyasla daha çok sayıda sözcük içeren bir Çince sözlük ile başlanmıştır. İlk evrede sınıflandırılan belgeler, destek vektör makineleri sınıflandırıcısının eğitilmesi amacıyla eğitim seti olarak kullanılmaktadır. Bu test verisi daha sonra ilk evrede elde edilen sonuçların gözden geçirilmesi için kullanılmaktadır. Geliştirilen yöntem, on farklı alandaki değerlendirme verileri üzerinde test edilmiş ve daha önceki çalışmalara kıyasla daha yüksek başarım elde edildiği görülmüştür.

Rothfels ve Tibshirani [48] çalışmalarında, İngilizce film değerlendirmeleri üzerinde otomatik görüş sınıflandırması gerçekleştirmek için öğreticisiz bir yöntem geliştirmiştir. Bu çalışmada, [46] tarafından Çince ürün değerlendirmeleri üzerinde uygulanan yöntem, İngilizceye uyarlanmış ve film değerlendirmeleri için kullanılmıştır. Çincede etkin sonuçların elde edildiği yöntemin, İngilizce film değerlendirmelerinde yeterli başarım oranını elde edemediği görülmüştür. Bu, iki dilin farklı özelliklerinden ve alanlar arası farklılıktan kaynaklanmaktadır. Başarımın artırılabilmesi için Turney ve Littman [43] tarafından önerilen noktasal karşılıklı bilgiye dayalı duygu yönü belirleme ile sıfat öbeklerinin kutupları belirlenip kademeli olarak yeniden eğitime tabi tutulduğunda %50.3 olan doğru sınıflandırma oranının %65.5’e çıktığı gözlenmiştir. Tablo 3’te bu bölümde verilen öğreticisiz öğrenme yöntemlerine ilişkin özet bilgiler sunulmaktadır.

Tablo 3: Öğreticisiz öğrenme yöntemlerinin karşılaştırılması.

Referans	Yıl	Yöntemler	Veri Alanı	Değerlendirme Ölçütü	Değer
[43]	2002	Belgede geçen belirteç ve sıfatlara dayalı yön belirleme, Noktasal Karşılıklı Bilgi, Gizli Anlamsal Çözümleme	General Inquirer Lexicon	Doğru Sınıflandırma	%80.00
[44]	2008	WordNet sözcük veritabanı, sözlük-tabanlı sınıflandırma	Film, Haber, Blog Değerlendirmeleri	Doğru Sınıflandırma	%78.00
[45]	2008	Öğreticisiz öğrenme, Kademeli olarak yeniden eğitime dayalı sözlük geliştirme	Çince Görüş Değerlendirmeleri	F-ölçütü	%87.00
[46]	2008	Öğreticisiz öğrenme, Otomatik sözcük seçimi, Sezgisel bilgi, Tekrarlamalalı yeniden eğitim	Çince Görüş Değerlendirmeleri	F-ölçütü	%92.00
[47]	2009	Derlem, Görüş Sözlüğü, Öğreticisiz Öğrenme	Görüş Değerlendirmeleri	F-ölçütü	%89.35
[48]	2015	Öğreticisiz öğrenme	Film Değerlendirmeleri	Doğru Sınıflandırma	%64.50

5 Sonuçlar

Görüş madenciliğinde öğreticili öğrenme yöntemleri, yeterince büyük veri setleri ile eğitildiğinde yüksek doğru sınıflandırma oranları elde edilebilmektedir. Bunun yanı sıra, görüş madenciliğinde sınıflandırıcı toplulukları kullanılarak, farklı sınıflandırıcıların güçlü yönlerini taşıyan ve yüksek başarım elde eden yöntemler de geliştirilmektedir. Ancak, öğreticili öğrenme yöntemlerinin alan bağımlılığı yüksektir. Belirli bir alana ait veri seti ile eğitilerek bir sınıflandırma modeli oluşturan sınıflandırıcının, başka bir alana ait veri seti aracılığıyla test edildiğinde performansının önemli ölçüde düştüğü gözlenmektedir. Alan bağımlılığına ek olarak, belirli bir alana ait özellikler bile zamanla değişmektedir. Belirli bir alana özgü derleme görüş bildiren etiketlerin eklenmesi önemli bir maliyet oluşturmaktadır. Bunun yanı sıra, eğitim verisi ile test verisi arasındaki dağılımlar farklılaştığında, öğreticili öğrenme yöntemleri ile etkin sonuçlar alınamamaktadır. Bu nedenle, görüş madenciliği alanındaki çalışmaların önemli bir bölümü, yarı-öğreticili ya da öğreticisiz öğrenme yöntemlerine dayalı bir yöntem geliştirilmesine odaklanmaktadır. Yarı-öğreticili öğrenme yöntemlerinde, az sayıda etiketli veri ile büyük miktarda etiketsiz veri eğitim için kullanılmaktadır. Yarı-öğreticili öğrenme yöntemleri, görüş sınıflandırmada karşılaşılan iki temel problemin ortadan kaldırılması için sıklıkla kullanılmaktadır. Bu problemlerden birincisi, bir alanda eğitilen bir sınıflandırıcının başka bir alanda kullanılmasına ilişkin alan-transfer problemidir. Diğer problem ise farklı dillerde yapılan görüş sınıflandırmadır. Öğreticisiz öğrenmeye dayalı yöntemler ise alan bağımlılığı problemini ortadan kaldırmaktadır. Öğreticisiz öğrenme yöntemleri kullanılarak görüş sınıflandırma yapılırken, alan bağımsız görüş sözlükleri kullanılmaktadır. Genel olarak değerlendirildiğinde, doğru sınıflandırma başarımı açısından en yüksek sonuçların öğreticili öğrenme yöntemleri ile elde edildiği, öğreticisiz ya da yarı-öğreticili öğrenmeye dayalı yöntemlerin doğru sınıflandırma oranlarının kısmen daha düşük olduğu görülmektedir. Bilişim teknolojilerindeki ilerlemeler ile birlikte Web, kullanıcıların birçok farklı alandaki bilgi ve görüşlerini etkin bir biçimde paylaşabilecekleri sosyal bir ortam haline gelmiştir. Web, alan ve içerik bakımından zengin bir ortamdır. Sosyal çevrimiçi veri, heterojen, az yapılandırılmış ve hızlı bir biçimde değişen bir yapıya sahiptir. Sosyal veri hızlı bir oran ile artmakta; kullanışlı ve yararlı verilerin yanı sıra; birçok kullanışsız ve gereksiz bilgi de içermektedir. Bu durumlar göz önünde bulundurulduğunda, alan bağımlılığı düşük ve büyük veriler üzerinde etkin bir biçimde çalışabilen görüş sınıflandırma yöntemlerinin geliştirilmesi önemli bir araştırma alanı konumuna gelmektedir. Bu doğrultuda, görüş madenciliği alanında, sınıflandırma modelinin oluşturulmasında kullanılacak eğitim setindeki örneklerin etkin bir biçimde seçilmesi, etkin öznitelik seçim yöntemleri geliştirilmesi ve yarı-denetimli ya da denetimsiz etkin öğrenme modellerinin oluşturulması önemli araştırma konuları haline gelmektedir.

6 Kaynaklar

[1] Ganesan K, Kim H. D. "Opinion Mining Tutorial (Sentiment Analysis)".http://www.slideshare.net/KavitaGanesan/opinion-mining-kavitahyunduk00 (15.01.2015).

[2] Medhat W, Hassan A, Korashy H. "Sentiment analysis algorithms and applications: a survey". *Ain Shams Engineering Journal,* 5(4), 1093-1113, 2014

[3] Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. "Lexicon-Based methods for sentiment analysis". *Computational Linguistics*, 37(2), 267-307, 2011.

[4] Pang B, Lee L, Vaithyanathan S. "Thumbs up?: Sentiment classification using machine learning techniques". *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, PA, USA, 6-7 July 2002.

[5] Pang B, Lee L. "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts". *42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain, 21-26 July 2004.

[6] 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.

[7] 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.

[8] McDonald R, Hannan K, Neylon T, Wells M, Reynar J. "Structured models for fine-to-coarse sentiment analysis". *45th Annual Meeting of the Association of Computational Linguistics,* Prague, Czech Republic, 23-30 June 2007.

[9] Zaidan OF, Eisner J, Piatko CD. "Using 'annotator rationales' to improve machine learning for text categorization". *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, Rochester, NY, USA, 22-27 April 2007.

[10] Tan S, Zhang J. "An empirical study of sentiment analysis for Chinese document". *Expert Systems with Applications*, 34(4), 2622-2629, 2008.
[11] Prabowo R, Thelwall M. "Sentiment Analysis: a combined approach". *Journal of Informetrics,* 3(2), 143-157, 2009.
[12] 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.
[13] 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.
[14] Zhao YY, Qin B, Liu T. "Integrating intra-and inter-document evidences for improving sentence sentiment classification". *Acta Automatica Sinica*, 36(10), 1417-1425, 2010.
[15] Bai X. "Predicting consumer sentiments from online text". *Decision Support Systems,* 50(4), 732-742, 2011.
[16] Chen CC, Tseng YD. "Quality evaluation of product reviews using an information quality framework". *Decision Support Systems,* 50(4), 755-768, 2011.
[17] Wang S, Li D, Song X, We, Y, Li H. "A feature selection method based on improved fisher's discriminant ration for text sentiment classification". *Expert Systems with Applications,* 38(7), 8696-8702, 2011.
[18] Xia R, Zong C, Li S. "Ensemble of feature sets and classification algorithms". *Information Sciences,* 181(6), 1138-1152, 2011.
[19] Kang H, Yoo SJ, Han M. "Senti-Lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews". *Expert Systems with Applications,* 39(5), 6000-6010, 2012.
[20] Sun Y, Wong AKC, Kamel SM. "Classification of imbalanced data: a review". *International Journal of Pattern Recognition and Artificial Intelligence,* 23(4), 687-719, 2009.
[21] Li YM, Li TY. "Deriving Market intelligence from microblogs". *Decision Support Systems,* 55(1), 206-217, 2013.
[22] Moraes R, Valiati JF, Neto WPG. "Document-Level sentiment classification: an empirical comparison between SVM and ANN". *Expert Systems with Applications,* 40(2), 621-633, 2013.
[23] Wang G, Sun J, Ma J, Xu K, Gu J. "Sentiment classification: the contribution of ensemble learning". *Decision Support Systems,* 57, 77-93, 2014.
[24] 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.
[25] 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.
[26] Lin C. Probabilistic Topic Models for Sentiment Analysis on the Web. PhD Thesis, University of Exeter, Exeter, UK, 2011.
[27] 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.
[28] 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.
[29] 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.
[30] 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.
[31] 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.
[32] 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.
[33] Li T, Zhang Y, Sindhwani V. "A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge". *47th Annual Meeting of the Association for Computational Linguistics (ACL)*, Suntec, Singapore, 2-7 August 2009.
[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] Li S, Huang CR, Zhou G, Lee SYM. "Employing personal/impersonal views in supervised and semi-supervised sentiment classification". *48th Annual Meeting of the Association for Computational Linguistics (ACL),* Uppsala, Sweden, 11-16 July 2010.
[37] He Y, Zhou D. "Self-Training from labelled features for sentiment analysis". *Information Processing and Management*, 47(4), 606-616, 2011.
[38] He Y, Lin C, Alani H. "Automatically extracting polarity-bearing topics for cross-domain sentiment classification". *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, Portland, OR, USA, 19-24 June 2011.
[39] Hernandez OJ, Rodriguez JD, Alzate, L, Lucania M, Inza I, Lozano JA. "Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers". *Neurocomputing*, 92, 98-115, 2012.
[40] Hajmohammadi MS, Ibrahim R, Selamat A. "Bi-View semi-supervised active learning for cross-lingual sentiment classification". *Information Processing and Management*, 50(5), 718-732, 2014.
[41] Hajmohammadi MS, 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.

[42] Hajmohammadi MS, 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.

[43] Turney P, Littman M. "Unsupervised learning of semantic orientation from a hundred-billion-word corpus". Institute for Information Technology, National Research Council, Ottawa, Ontario, Canada, Technical report, *ERB-1094*, 2002.

[44] Andreevskai A, Bergler S. "When specialists and generalists work together: overcoming domain dependence in sentiment tagging". *46th Annual Meeting of the Association for Computational Linguistics (ACL)*, Columbus, OH, USA, 19-20 June 2008.

[45] Zagibalov T, Carroll J. "Unsupervised classification of sentiment and objectivity in Chinese text". *International Joint Conference on Natural Language Processing (IJCNLP)*, Hyderabad, India, 7-12 January 2008.

[46] Zagibalov T, Carroll J. "Automatic seed word selection for unsupervised sentiment classification of Chinese text". *22nd International Conference on Computational Linguistics (COLING'08)*, Manchester, UK, 18-22 August 2008.

[47] Qui L, Zhang W, Hu C, Zhao K. "SELC: a self-supervised model for sentiment classification". *18th Association for Computing Machinery conference on Information and Knowledge Management (ACM-CIKM)*, Hong Kong, China, 2-6 November 2009.

[48] Rothfels J, Tibsirani J. "Unsupervised Sentiment Classification of English Movie Reviews Using Automatic Selection of Positive and Negative Sentiment Items". http://nlp.stanford.edu/courses/cs224n/2010/reports/rothfels-jtibs.pdf (15.01.2015).