Eğiticili ve Geleneksel Terim Ağırlıklandırma Yöntemleriyle Duygu Analizi

Supervised and Traditional Term Weighting Methods for Sentiment Analysis

Mahmut Çetin
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
cetinmahmut@msn.com

M. Fatih Amasyalı
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
mfatih@ce.yildiz.edu.tr

***Özetçe*— Duygu analizi bir metin sınıflandırma problemi olup popülerliği ve ticari getirileri sebebiyle günümüzde üzerinde çokça çalışılan bir konudur. Metin sınıflandırmadaki en önemli nokta metinlerin nasıl temsil edilmesi gerektiğidir. Geleneksel eğiticisiz yöntemler yerine terimlerin sınıf dağılımlarını da hesaba katan eğiticili yöntemler literatürde sıklıkla kullanılmaya başlanmıştır. Bu çalışmada Türkçe Twitter gönderilerinden oluşan 2 veri kümesi üzerinde bu yöntemler çeşitli boyutlarda karşılaştırılmıştır. Sonuç olarak eğiticili yöntemlerin daha başarılı ve daha uygulanabilir oldukları görülmüştür.**

Anahtar Kelimeler; Duygu analizi, sentiment analizi, metin sınıflandırma, terim ağırlıklandırma, örüntü tanıma, makine öğrenmesi.

***Abstract*—Sentiment analysis is a text classifying problem and because of its popularity and commercial revenue, it has been widely studied. The most important point in text categorization is how to represent the texts. Instead of traditional methods, supervised term weighting methods which include terms' distribution of classes has been started to be used. In this study, these methods are compared in different dimensions on two datasets which consist Turkish Twitter posts. In conclusion, supervised term weighting methods are found more successful and applicable.**

Keywords; sentiment analysis, text classification, term weighting methods, pattern recognitio, machine learning.

I. GİRİŞ

"X'in Y filmindeki performansı harika", "X'in fiyatları çok yüksek", "sonunda benimde X'im var" vb. sosyal medya mesajlarının analizi gün geçtikçe daha çok firmanın ilgisini çekmektedir. Bir filme gitmeden önce kullanıcı yorumlarını okumak, bir telefonu almadan önce onu kullananların fikirlerini araştırmak günümüzde ticaretinin genellikle ilk aşamaları haline gelmiştir. Firmaların reklamları kadar diğer kullanıcıların yorumları bir ürün hakkındaki algımızı şekillendirmektedir. Bu önceden de böyleydi belki ama artık diğer kullanıcıların yorumlarına erişmek çok daha kolay olduğu için etkisi daha da güçlüdür. Bu durum ticari doğal dil işleme çalışmalarına olan ilgiyi arttırmıştır. Firmalar kendi ürünlerinden bahsedilen sosyal medya mesajlarından kamuoyu algılarını ölçmek istemektedirler. Bununla birlikte sosyal medyadaki veri çok büyük olduğundan bu işlemin elle yapılması oldukça güçtür. Bunun yerine bir firmadan ya da ürününden bahseden sosyal medya mesajının olumlu ya da olumsuz yargı içerdiğinin otomatik olarak bulunması fikri doğmuştur. Bu problem, bir doğal dil işleme problemi olarak formüle edilirse gelen bir mesajın hangi sınıfa ait olduğunun bulunması haline gelir ki bu bir metin sınıflandırma uygulamasıdır ve doğal dil işleme literatürü bu konuda oldukça geniş bir çözümler havuzuna sahiptir.

Metin sınıflandırma 2 temel alt probleme indirgenebilir. İlki metinlerin nasıl temsil edileceği, ikincisi ise hangi algoritma ile sınıflandırma yapılacağıdır.

Sosyal medya mesajlarından üzerinde en çok çalışılanı popülerliği, çeşitliliği ve erişim kolaylığı nedeniyle Tweet'lerdir. Sahip olduğu bu avantajların yanında 140 karakterler kısıtlanmış olması, kendine ait bir jargonunun olması, yazım hatalarının çok fazla olması gibi doğal dil işleme yöntemlerini zorlayan yönleri de bulunmaktadır. 140 karakterle kısıtlanmış olması, mesajdan elde edilebilecek veri miktarını azaltmakta, kendine ait bir jargonunun olması ve yazım hatalarının çok olması ise morfolojik analizi zorlaştırmaktadır.

Literatürde metinleri temsil etmek için en çok kullanılan yöntemler kelimelerin köklerinin, karakter ngramların metinlerdeki geçiş sayıları ve bunların ağırlıklandırılmış halleridir. Kelime köklerinin kullanılabilmesi için bir morfolojik çözümleyiciye ihtiyaç varken, karakter ngramları doğrudan kullanılabilmektedir.

Geleneksel yöntemlerde kullanılan terim frekansı (tf), terimin o metinde kaç kere geçtiğini ifade eder. Çok sıklıkla kullanılan terimlerin etkisinin azaltılması için terim frekansının ters doküman frekansı ile ağırlıklandırılmış hali (tfidf) kullanılmaktadır.

Eğiticili terim ağırlıklandırma yöntemlerinde ise terimlerin sınıflarda geçme dağılımları da işleme katılır. Literatürde bu

978-1-4673-5563-6/13/$31.00 ©2013 IEEE

konuda popüler olarak kullanılan 6 yöntem bulunmaktadır [1, 2, 3, 4].

Bu çalışmada Türkçe metinlerde duygu analizinde geleneksel 2 yöntemin ve eğiticili 6 yöntemin performansları 5 algoritma kullanılarak 2 veri kümesi üzerinde karşılaştırılmıştır.

Bildirinin 2. bölümünde eğiticili terim ağırlıklandırma yöntemleri 3. bölümde ise geleneksel yöntemler tanıtılmaktadır. 4. Bölümde kullanılan veri kümeleri verilmiştir. Deneysel sonuçlar 5. Bölümde yer almaktadır. 6. bölümde ise sonuçlar ve gelecek çalışma planlarımız verilmiştir.

II. EĞITICILI TERIM AĞIRLIKLANDIRMA YÖNTEMLERI

Literatürdeki eğiticili terim ağırlıklandırma yöntemleri iki sınıflı problemlere göre formüle edilmiştir [059]. Çalışmamızda bu yöntemleri çoklu sınıf problemlerine de uygulamak için her sınıf için bir özellik üretilmesi düşünülmüş ve (+) o andaki üretilen özelliğin sınıfını temsil ederken (-) de diğer bütün sınıfları temsil etmektedir. Çalışmamızda kullanılan 6 farklı ağırlıklandırma formülü aşağıda verilmiştir [1]:

$$tf\ x\ rf = tf\ x \log_2 \left(2 + \frac{n^+(x)}{\max(1, n^-(x))}\right) \quad (1)$$

$$tf\ x\ KL = tf\ x \left(\frac{n^+(x)}{N^+ + N^-}\right) \left|\frac{n^+(x)}{n^-(x)}\right| \quad (2)$$

$$tf\ x\ \Delta_1 = tf\ x\ \log_2 \frac{N^+\ x\ n^-(x)}{N^-\ x\ n^+(x)} \quad (3)$$

$$tf\ x\ \Delta_2 = tf\ x\ \log \frac{(n^+(x) - N^+ + 0.5)\ x\ N^- + 0.5}{(n^-(x) - N^- + 0.5)\ x\ N^+ + 0.5} \quad (4)$$

$$tf\ x\ F_1 = tf\ x\ \log \frac{(n^+(x) + kN^- + k)}{(n^-(x) + kN^+ + k)} \quad (5)$$

$$tf\ x\ F_2 = tf\ x\ \log \frac{(n^+(x) + kN^- - n^- + k)}{(n^-(x) + kN^+ - n^+ + k)} \quad (6)$$

Formüllerde görülen $n^+(x)$ herhangi bir terimin ya da karakter gramın o sınıfta kaç metinde geçtiğinin sayısını temsil ederken $n^-(x)$ diğer bütün sınıflardaki kaç adet metinde geçtiğini temsil etmektedir. N^+ o sınıftaki metinlerin sayısını verirken N^- diğer bütün sınıflardaki metinlerin sayısını temsil etmektedir. *F1* ve *F2* yöntemlerindeki k ise sabit bir sayı olup 1 ya da 2 olarak belirlenebilmektedir. Bizim bu çalışmamızda 1 olarak kullanılmıştır [3].

Çalışmanın ilerleyen bölümlerinde buradaki formüllerle ifade edilen yöntemler şu kısaltmalarla anılacaktır: RF (1. Eşitlik), KL (2. Eşitlik), D1 (3. Eşitlik), D2 (4. Eşitlik), F1 (5. Eşitlik), F2 (6. Eşitlik).

III. GELENEKSEL TERİM AĞIRLIKLANDIRMA YÖNTEMLERİ

Bu bölümde sıklıkla kullanılan 2 yöntemden bahsedilmiştir.

Terim Frekansları: Bu temsil yönteminde metinler içerdikleri terimlerin frekanslarıyla ifade edilir. Bu terimler kelimelerin kendileri, kökleri ya da karakter gramlar olarak belirlenebilir. Bu yönteme göre satırlarında metinlerin, sütunlarında terimlerin yer aldığı bir matris oluşturulur. Matrisin [i, j] gözünde i. Metinde j. kelimenin kaç kere geçtiği bilgisi tutulur. Matrisin satır sayısı metin sayısına, sütun sayısı ise tüme metinlerde geçen farklı kelimelerin sayısına eşittir. Terim olarak kelime kökleri kullanıldığında kelimelerin morfolojik çözümlemesi için Zemberek [5] aracı kullanılmıştır.

Ters Doküman Frekansı: Terim frekanslarından farklı olarak terimlerin geçiş sayıları aşağıdaki formüldeki gibi ağırlıklandırılmaktadır [2].

$$tf\ x\ idf = tf\ x\ \log \frac{n}{n^+(x) + n^-(x)} \quad (7)$$

Formül 7'de n veri setinin içerdiği metin sayısını, $n^+(x) + n^-(x)$ ise x terimini içeren toplam metin göstermektedir. Her metinde çoğunlukla geçen terimler için log'un içi 1'e log'un sonucu da 0'a yaklaşarak o terimin ağırlığını düşürmektedir.

$$\alpha + \beta = \gamma \quad (1) \quad (1)$$

IV. KULLANILAN VERİ KÜMELERİ

Çalışmalarımızda kullanılan iki ayrı veri kümesi bulunmaktadır. Bunlardan ilki (A veri kümesi) telekom sektöründeki özel bir şirkete ait Twitter gönderilerinden oluşmaktayken, diğeri de (B veri kümesi) aynı sektördeki bir diğer şirket hakkındaki gönderilerden oluşmaktadır. Her iki veri kümesinde de 6000'er örnek bulunmaktadır. Bu kümeler olumlu, olumsuz ve nötr olmak üzere 3 ayrı sınıfa el ile ayrılmıştır. A veri kümesinde pozitif sınıfta 3040, negatif sınıfta 1847 ve nötr sınıfta 1113 adet gönderi bulunmakta iken, B veri setinde pozitif sınıfta 2680, negatif sınıfta 2316 ve nötr sınıfta ise 1004 adet gönderi bulunmaktadır.

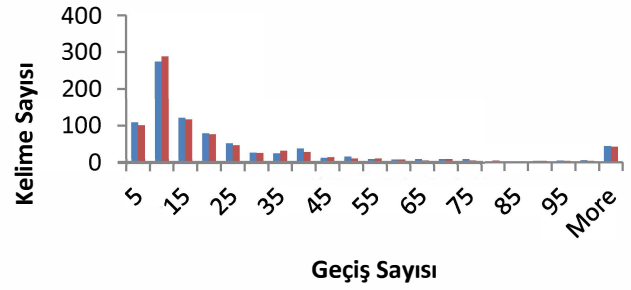
Daha sonra sınıf dağılımları eşit olacak şekilde veri kümeleri eşit boyutlu eğitim ve test kümelerine ayrılmıştır.

V. DENEYSEL SONUÇLAR

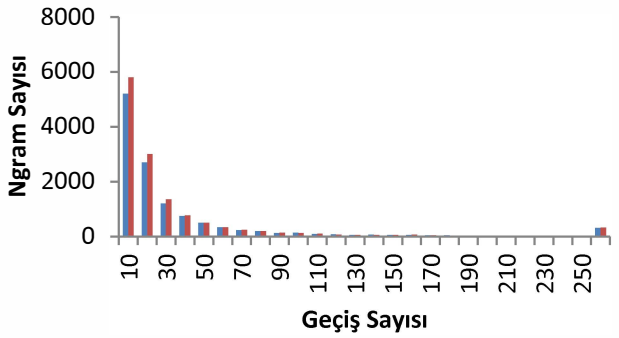
4. bölümde anlatılan metinler 2. ve 3. bölümde anlatılan yöntemler ile ifade edildikten sonra WEKA[12] kütüphanesinde yer alan Naive Bayes (NB), Karar Destek Makinesi (SMO), Karar Ağacı (J48), Rastgele Orman (RF) ve 1 en yakın komşu (IB1) sınıflandırma algoritmalarıyla model oluşturulmuştur. Test setinin içerdiği metinler de temsil edildikten sonra test seti üzerindeki başarılar ölçülmüştür.

Kullanılan temsil yöntemlerinin kelime kökleri ve karakter gramları kullandığını daha önceden belirtmiştik. Fakat hemen her metinde geçen terimlerin ayırt ediciliği az iken, çok az sayıda geçen terimlerde gereksiz yere veri kümesinin büyümesine sebep olmaktadır. Bu nedenle terim sayısını azaltmak için hem karakter gramlara hem de kelime köklerine bir frekans filtreleme uygulanarak 5-1000 aralığında geçen terimler değerlendirmeye alınmıştır. Şekil 1ve 2'de geçiş sayılarına göre terimlerin sayıları gösterilmiştir. Görüldüğü

gibi kelimeler ve karakter gramlar daha çok düşük frekanslara sahiplerdir. Mavi renkli sütunlar A veri kümesini kırmızı renkli sütunlar B veri kümesini temsil etmektedir.



ŞEKİL I. KELİMELERİN GEÇİŞ SAYILARINA GORE DAĞILIMLARI



ŞEKİL II. KARAKTER NGRAMLARIN SAYILARINA GÖRE DAĞILIMLARI

Tablo 1, 2, 3 ve 4'te A ve B veri kümeleri üzerinde terim olarak kelime köklerinin ve karakter ngramları (2,3, ve 4 ngram) kullanıldığında, 5 algoritmanın 8 terim ağırlıklandırma yöntemi ile elde edilen sonuçları verilmiştir. Ağırlıklandırma yöntemlerinin ve algoritmaların genel başarılarını değerlendirmek için satır ve sutunların ortalamaları alınmıştır.

TABLO I. A VERI KÜMESINDE KELIME KÖKLERININ BAŞARILARI (%)

Metot	NB	SMO	IB1	J48	RF	Ort.
TF	53,7	59,9	49,1	53,7	57,5	54,8
TFIDF	53,7	59,9	49,1	53,7	57,5	54,8
D1	62,2	64,2	57,6	62	60	61,2
D2	38,6	50,6	50,9	60,2	54,6	51,0
F1	62,2	**63,7**	57,1	61,6	59,8	60,9
F2	62,1	63,3	57,8	62,5	61,9	**61,5**
RF	62,1	63,2	58,4	61,7	60,0	61,1
KL	52,2	52,9	50,2	51,4	53,4	52,0
Ort.	55,9	**59,7**	53,8	58,4	58,1	

TABLO II. A VERI KÜMESINDE KARAKTER NGRAMLARIN BAŞARILARI (%)

Metot	NB	SMO	IB1	J48	RF	Ort.
TF	51,7	61,3	39,2	51	56,7	52,0
TFIDF	51,7	61,3	39,2	51	56,7	52,0
D1	61,3	64,7	59,1	60,5	61,5	61,4
D2	48,6	60,5	52,2	55,2	57,9	54,9
F1	61,1	64,9	59,5	60,8	63,5	**62,0**
F2	60,4	64,8	59,1	61,1	62,5	61,6
RF	60,3	**65,5**	58,9	60,9	61,3	61,4
KL	43,3	52,6	50,8	51,8	54,7	50,6
Ort.	54,8	**62,0**	52,3	56,5	59,4	

Tablo 1 ve 2 incelendiğinde TF ve TFIDF arasında bir fark bulunmamaktadır. Buna, sıklıkla kullanılan kelimelerin oranının çok az olması sebep olmuştur. Buradan terim doküman matrisinin çok seyrek olduğu da öngörülebilir.

TABLO III. B VERI KÜMESINDE KELIME KÖKLERININ BAŞARILARI (%)

Metot	NB	SMO	IB1	J48	RF	Ort.
TF	54,3	60,2	51,3	53,2	59,2	55,6
TFIDF	54,3	60,2	51,3	53,2	59,2	55,6
D1	61,7	62,6	56	61	58,5	60,0
D2	50,2	57,7	48,9	59,7	54,1	54,1
F1	61,5	62,6	57	61,7	60,1	**60,6**
F2	61,5	**62,8**	57,2	59,6	59,8	60,2
RF	61,2	61,5	57,5	59,8	59,4	59,9
KL	51,8	50,8	49,6	48,5	50,1	50,2
Ort.	57,1	**59,8**	53,6	57,1	57,6	

TABLO IV. B VERI KÜMESINDE KARAKTER NGRAMLARIN BAŞARILARI (%)

Metot	NB	SMO	IB1	J48	RF	Ort.
TF	51,2	62,9	49,4	49,6	56,8	54,0
TFIDF	50,1	58,4	49,4	49,6	56,8	52,9
D1	60,7	64,7	60,4	61,9	63,1	62,2
D2	49,1	60,6	51,9	57,5	58,8	55,6
F1	60,5	64,8	61,2	62,6	62,4	**62,3**
F2	60,1	64,7	61	61,2	61,9	61,8
RF	60,8	**65,3**	59,9	62,6	62,2	62,2
KL	41,8	51,2	50,7	53,5	53,7	50,2
Ort.	54,3	**61,6**	55,5	57,3	59,5	

2 veri kümesi üzerinde yapılan denemelerin sonuçları Tablo 5'te özetlenmiştir.

TABLO V. VERI KÜMELERINDEKI SONUÇLAR

		En başarılı alg. - terim ağırlıklandırma ikilisi ve başarısı	En başarılı / başarısız algoritma	En başarılı / başarısız terim ağırlıklandırma yöntemi
A veri kümesi	Kelime kökleri	SMO-F1-63,7	SMO / IB1	F2 / D2
	Karakter Ngramları	SMO-RF-65,5	SMO / IB1	F1 / KL
B veri kümesi	Kelime kökleri	SMO-F2-62,8	SMO / IB1	F1 / KL
	Karakter Ngramları	SMO-RF-65,3	SMO / NB	F1 / KL

Tablo 5 incelendiğinde her iki veri kümesinde de, ister kelime kökleri isterse karakter ngramları kullanılsın en başarılı terim ağırlıklandırma yöntemleri hep eğiticililer olmuştur. Ancak eğiticililerden KL ve D2 geleneksel yöntemlerden daha başarısızdırlar. En başarılı algoritma her zaman SMO olmuştur. En kötü algoritma ise genelde IB1 olmuştur. Ayrıca her iki veri kümesi için de karakter ngramlarıyla kelime köklerine göre daha yüksek başarıların elde edildiği görülmektedir.

Eğitilici ağırlıklandırma yöntemlerinin başarılarının yanında bir diğer avantajları ise metin temsilindeki özellik sayılarının azlığıdır. Tablo 6'da A ve B veri kümelerinde metinlerin temsilinde kullanılan her bir yöntem için özellik sayıları gösterilmiştir.

TABLO VI. VERI KÜMELERINDE YÖNTEMLERIN OLUŞTURDUĞU ÖZELLIK SAYILARI

		TF- TFIDF için özellik sayısı	D1-D2-F1-F2-RF-KL için özellik sayısı
A veri kümesi	Kelime kökleri	848	3
	Karakter Ngramları	12326	9
B veri kümesi	Kelime kökleri	824	3
	Karakter Ngramları	13478	9

Eğiticili yöntemlerde eğitim kümelerindeki tüm terimlerin her sınıf için ağırlıkları hesaplandıktan sonra test örnekleri sayısallaştırılırken, içerdikleri terimlerin ağırlıklarının ortalaması alınarak sınıf sayısı adet özellikle temsil edilmektedirler. Bu sebeple kelime köklerinin eğiticili yöntemlerdeki özellik sayıları sınıf sayısına yani 3'e eşittir. Karakter ngramlarında ise 2, 3, ve 4 ngramların her biri için bu işlem tekrar edildiğinden test örnekleri 3*3 özellikle temsil edilmişlerdir. Eğiticisiz yöntemlerde ise metinlerin boyutları eğitim kümesinin içerdiği tekil terim sayısı kadardır ki bu rakam Tablo 6'da da görüldüğü gibi çok yüksektir. Yüksek boyutlu veriler, sınıflandırma algoritmalarının ürettikleri modellerin karmaşıklığını ve dolayısıyla çalışma zamanını arttırmaktadır. Binlerce boyuta sahip bir veri kümesine göre sınıf sayısı kadar boyuta sahip bir veri kümesi çok daha hızlı modellenebilmektedir.

VI. SONUÇ

Metin sınıflandırmanın günümüzdeki en popüler ve ticari uygulaması olan duygu analizinde kullanıcıların sosyal medya mesajlarının olumlu ya da olumsuz yargı içerdiğinin otomatik olarak bulunması amaçlanmaktadır.

Bu çalışmada metin sınıflandırmada kullanılan eğiticili ve eğiticisiz terim ağırlıklandırma yöntemlerinin iki adet Türkçe Twitter gönderimi veri kümesi üzerinde karşılaştırmalı analizi yapılmıştır. Karşılaştırmada metinlerin kelime kökleri ve karakter ngramlarıyla temsili, çeşitli sınıflandırma algoritmalarının performansları incelenmiştir. Her iki veri kümesinde de paralel olarak elde ettiğimiz sonuçlar aşağıda verilmiştir:

-Metin temsilinde karakter ngramlarının, kelime köklerine göre daha başarılıdır.

-Terim ağırlıklandırmada eğiticili yöntemlerin geleneksel eğiticisiz yöntemlere göre daha başarılı ve içerdikleri az sayıda özellik sayısı sebebiyle daha kolay ve hızlı modellenebilmektedir. Bu sebeple özellik sayısının geleneksel yöntemler kullanıldığında binlerce olduğu veri kümelerinde eğiticili yöntemlerin uygulanabilirliği daha fazladır.

-Terim ağırlıklandırmada eğiticili D2 ve KL haricindeki tüm yöntemler eğiticisiz yöntemlerden daha başarılıdır. Elde ettiğimiz sonuçlar literatürde İngilizce için yapılmış çalışmalarla KL'nin düşük başarılı sonuç vermesi dışında paralellik göstermiştir. Bunun sebebi 2 sınıflı problemler dışında, terimlerin bir sınıftaki geçiş sayısının diğer sınıflardaki geçiş sayılarına oranının farklılık gösterememesi, az da olsa gösterenlerin de yapılan çarpma işlemi ile etkisinin azaltılmasıdır.

-Denenen algoritmalar arasında SMO en yüksek sınıflandırma başarısına sahiptir.

Elde ettiğimiz sonuçlar literatürde İngilizce için yapılmış çalışmalarla paralellik göstermektedir.

Kelime köklerinin karakter ngramlarına göre daha başarısız olmalarının sebebi yanlış yazımlar ve sözlük dışı kelimelerin fazlalığından oluşan morfolojik çözümleme problemleridir. Birçok kelime çözümlenemediğinden metinlerin temsilinde yer alamamaktadır. Gelecek bir çalışma olarak, bu problemi çözmek için sosyal medya mesajlarına özel olarak geliştirilen bir otomatik imla düzeltici ile ön işleme yapılması planlanmaktadır.

Bu projedeki çalışmalar Ericsson Türkiye tarafından desteklenmektedir.

KAYNAKÇA

[1] Tam T. Nguyem, Kuiyu Chang, Siu Cheung Hui(2011), "Supervised Term Weighting for Sentiment Analysis".

[2] Man Lan, Chew Lim Tan(2007), Supervised and Traditional Term Weighting Methods for Automatic Text Categorization.

[3] Tam T.Nguyen, Kuiyu Chang, Siu Cheung Hui, Probabilistic Supervised Term Weighting for Binary Text Categorization

[4] Justin Martineau and Tim Finin(2009), Delta TFIDF: An Improved Feature Space for Sentiment Analysis.

[5] code.google.com/p/zemberek/