Duygu Analizinde Transfer Öğrenme

Ahmet NİŞLİ[1] M. Fatih AMASYALI[2]

[1,2]Bilgisayar Mühendisliği Bölümü
Elektrik-Elektronik Fakültesi
Yıldız Teknik Üniversitesi, Davutpaşa, İSTANBUL

Email: nisliahmet@gmail.com mfatih@ce.yildiz.edu.tr

Özet

Günümüzde firmalar kullanıcılarının görüşlerini sosyal medyadan takip etmektedir ve analiz etmek istemektedirler. Ancak bu verileri otomatik analiz edebilmek için etiketli verilere ihtiyaçları vardır. Ancak etiketsiz verileri etiketlemek zor ve maliyetli bir iştir. Bu çalışmada aynı sektörde bulunan iki firmadan birinin sahip olduğu etiketli verilerin diğer firma tarafından da kullanılabilir hale getirilebilmesi için çalışmalar yapılmıştır. Tweetleri analiz edebilmek için orijinal kelime, 3gram, kelime kökleri ve kelime kökleri ile kelime önerileri yöntemleri kullanılarak ve her bir yöntem ile binary ve tf (özellik frekansı) şeklinde veri setleri oluşturuldu. Analiz için SMO, Random Forest, J48, Ibk ve Naive Bayes sınıflandırıcıları kullanılmış ve transfer yöntemleri denenmiştir. Transfer öğrenme konusunda bütün modeller içinde en başarılı metin sınıflandırıcının orijinal kelimelerle oluşturulmuş binary dosyası, en başarılı sınıflandırıcının Random Forest ve en başarılı transfer yönteminin ise eş-zıt tweetler ekleme yöntemi olduğu görülmüştür.

1. Giriş

Günümüzde, sosyal medya çok yaygın bir şekilde kullanılmaktadır. Birçok insan sosyal medyadan görüşlerini bildirmektedir. Firmalar sosyal medya aracılığıyla kullanıcılarının kendileri hakkındaki görüşlerini öğrenebilirler ve onlarla iletişime geçebilirler. Şirketler, kullanıcıların kendileri hakkındaki görüşlerini analiz edebilmek için kendi sistemlerini kurmaktadır. Kullanıcı görüşlerinin analiz edilebilmesi için etiketli verilere ihtiyaç vardır. Fakat etiketsiz verileri etiketlemek çok maliyetli ve uzun süren bir iştir. Bu projede transfer öğrenme ile bir firma için olan veri setini aynı sektörde bulunan diğer firma için de kullanılabilir hale getirilmesi amaçlanmıştır. Bu işlem ile etiketleme maliyetinin azaltılması hedeflenmiş ve o firmanın analiz başarısının artırılması hedeflenmektedir.

Transfer öğrenme için Raina ve Koller [1], eğitim setindeki verilerdeki önemli kelimeleri bulmuş ve transfer edilecek veri setindeki verileri bu kelimelerin geçme durumuna göre transfer ederek başarıyı artırmışlardır. Gerekli etiketli veri kümesinin boyutunun azaltılması transfer öğrenmenin amacı olduğu gibi aktif öğrenmenin de amacıdır. Bu amaçla duygu durum analizi için Çetin ve Amasyalı [2], sınıflandırıcıları ve aktif öğrenme yöntemlerini kullanarak aynı performansı daha az eğitim verisi kullanarak almışlardır.

Bu çalışmada, firmalar arası veri transfer çalışmaları yapılmıştır. Metin sayısallaştırma yöntemlerinin veya sınıflandırma algoritmalarının transfer öğrenmede etkisi olup olmadığı sorularına cevap aranmıştır. Ayrıca çeşitli transfer öğrenme yöntemleri denenmiş ve sonuçları araştırılmıştır. Yazının ikinci bölümünde yöntemler açıklanmıştır. Üçüncü bölümde deneysel sonuçlar açıklanmıştır. Dördüncü bölümde sonuç, beşinci bölümde ise kaynakçaya yer verilmiştir.

2.Kullanılan Yöntemler

Çalışmamızda elde ettiğimiz tweetleri Weka'da analiz edebilmek amacıyla tweetlerden arff dosyası üretmemiz gereklidir. Arff dosyası üretirken veri tabanı içindeki noktalama işaretleri ve linkler kaldırılmıştır. Çalışmamızda arff dosyası oluştururken kullandığımız metin sayısallaştırma yöntemleri şunlardır:

Orijinal Kelime: Arff dosyamızda bulunan özellikler tweetlerde bulunan orijinal kelimelerdir.

3 gram: Arff dosyamızda bulunan özellikler tweetlerin üçer karakterlere ayrılmış halidir.

Kelime Kökleri: Arff dosyamızda bulunan özellikler orijinal kelimelerin kökleri şeklindedir. Bu yöntemde javada bulunan Zemberek kütüphanesi kullanılmıştır.
Kelime Kökleri ve Kelime Önerileri: Arff dosyamızda bulunan kelimelerin kökleri kullanılmış aynı zamanda Türkçe olmayan kelimelere Zemberek[5] yardımıyla kelime önerisinde bulunularak özellikler belirlenmiştir.

Her bir metin sayısallaştırma işlemi binary ve tf olarak hazırlanmıştır. Binary dosyasında kelimenin o tweette olup olmadığı (0-1), tf dosyasında ise kaç defa geçtiği ifade edilmektedir.

Çalışmamızda firmalar arası transfer yapmak için arff dosyası oluşturulurken hangi firmadan transfer yapmak istersek o firmanın ismini diğer firma ile değiştirmek gerekmektedir. Örneğin; Vodafone firmasından Turkcell'e transfer işlemlerinde Vodafone veri tabanında "Vodafone çok iyi" tweeti varsa bu tweet "Turkcell çok iyi" şekline çevrilerek veri seti oluşturulur.

Arff dosyalarını oluşturduktan sonra çalışmamızda veri setimizde transfer edilecek veri seti ek eğitim seti olarak adlandırılır. Transfer edilecek eğitim seti ise eğitim ve test seti olarak ikiye ayrılır.Eğitim setine, ek eğitim setinin tamamı transfer edilebileceği gibi çeşitli yöntemlerle bir kısmı da transfer edilebilir. Çalışmamızda çeşitli yöntemler denenmiştir. Bu yöntemler şunlardır:

Hepsinin Eklenmesi: Ek eğitim setindeki bütün verilerin eklenmesidir.
Nötr Olmayan: Sadece nötr etikete sahip olmayan veriler eklenmiştir. Örneğin; "Vodafone'a geçiyorum kullanan var mı?" tweeti analiz etmede çok yarar sağlamayacağından bu yöntemde eklenmemiştir.
Firma Adı Geçmeyen: Her iki firmanın adının geçmediği verilerin eklenmesidir. Örneğin; "Vodafone bizim evde cekmiyor tek ceken turkcell". Gibi her iki firmanın da adının geçtiği tweetler analiz edilirken çelişki oluşturabileceği için bu tweetler alınmamıştır.
Hem Nötr Olmayan Hem de Firma Adı Geçmeyen Ek Veri Seti Oluşturma Yöntemi: Elimizdeki eğitim setindeki kelimelerin eş zıt anlamları kullanılarak yeni tweetler oluşturma ve ek eğitim seti ile birlikte ana eğitim setinin üzerine eklenmesidir. Bu yöntemde ilişkisel sözlük olan Tdk-wiki ilişkisel sözlüğü kullanılmıştır[3]. Eş tweet örneğin: "Turkcell sınırsız internet paketi almayın" ise yeni oluşturulan tweet "Turkcell limitsiz internet paketi almayın". Zıt tweet örneğin:"Turkcell internet çok kötü" ve etiketi "olumsuz" ise yeni oluşturulacak tweet "Turkcell internet çok iyi" olup etikette "olumlu" hale çevrilir.
Bfs Kelime Analizi Yöntemi: Weka'daki Best First Search fonksiyonu ile elimizdeki eğitim seti analiz edilir ve sonucunda anahtar kelimeler bulunur. Çalışmamızda iki yöntemle kullanılmıştır. Birinci yöntemde eğer ek eğitim setindeki tweet içinde anahtar kelime varsa, ikinci yöntemde ise eğer ek eğitim setindeki tweet içindeki anahtar kelime sayısı ikiden az ise transfer edilmiştir.
Ki Kare Kelime Analiz Yöntemi: Weka'daki Ki Kare fonksiyonu ile elimizdeki eğitim seti analiz edilir ve sonucunda anahtar kelimeler bulunur. Çalışmamızda iki yöntemle kullanılmıştır. Birinci yöntemde eğer ek eğitim setindeki tweet içinde anahtar kelime varsa, ikinci yöntemde ise eğer ek eğitim setindeki tweet içindeki anahtar kelime sayısı ikiden az ise transfer edilmiştir.
Kosinüs Benzerliği Yöntemi: Bu yöntemde elimizdeki eğitim seti ve ek eğitim setindeki tweetlerin birbirlerine olan yakınlıkları kosinüs benzerliği Eşitlik 1'le hesaplanır.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

Kosinüs benzerliği formülünde "A" ana eğitim kümesinin tweetlerini "B" ise ek eğitim kümesindeki tweetleri temsil etmektedir. Her bir "A" her bir "B" ile karşılaştırılır. Bu işlem sonucu ortaya çıkan sayı 0 ile 1 arasındadır ve 1'e ne kadar yakınsa iki tweet o kadar benzerdir denilebilir. Bütün tweetlerin karşılaştırılarak yakınlık matrisi elde edilir.
Yakınlık matrisi elde edildikten sonra verileri transfer ederken üç yöntem kullanılmıştır. Birinci yöntemde en benzer %50 tweet alınarak transfer edilmiştir. İkinci yöntemde en az benzer %50 tweet alınarak transfer edilmiştir. Üçüncü yöntemde ise en az benzer %25 ve en benzer %25 transfer edilerek testler yapılmıştır.
Korelasyon Benzerliği Yöntemi: Bu yöntemde elimizdeki ana eğitim seti ve ek eğitim setindeki tweetlerin birbirleriyle benzerlikleri Weka'da bulunan korelasyon fonksiyonuyla hesaplanır. Eğitim seti ile eş eğitim setindeki tweetler karşılaştırılarak benzerlik matrisi elde edilir. Benzerlik matrisi elde edildikten sonra transfer aşamasında üç yöntem kullanılmıştır. Birinci yöntemde en ilişkili %50 tweet alınarak transfer edilmiştir. İkinci yöntemde en az ilişkili %50 tweet alınarak transfer edilmiştir. Üçüncü yöntemde ise en az

ilişkili %25 ve en ilişkili %25 transfer edilerek testler yapılmıştır.

3.Deneysel Sonuçlar

Çalışmamızda iki adet veri tabanı kullanılmıştır. Birinci veri tabanı 1500 Turkcell ve 1500 Vodafone tweetlerinden oluşmuştur. İkinci veri tabanı 2000 Turkcell ve 2000 Ttnet tweetlerinden oluşmuştur.

Tablo 1. Kullanılan Veri Setleri

Veri Seti	Şirket	Olumlu	Olumsuz	Nötr
Turkcell-Vodafone Veri Seti	Turkcell	319	1084	97
	Vodafone	450	975	75
Turkcell-Ttnet Veri Seti	Turkcell	685	717	598
	Ttnet	637	711	652

Çalışmamızda her bir metin sayısallaştırıcı tekniği ile binary ve tf olarak veri setleri üretilmiştir. Ayrıca veri setleri her iki firmanın birbirine transferleri için ayrı ayrı oluşturulmuştur.

Test sırasında ve transfer yönteminde yapılan işlemler şunlardır:

1. Veri setimizdeki transfer edilecek şirketin verilerine ek eğitim seti denilir. Test edeceğimiz şirketin verileri de önce karıştırılır sonra eğitim ve test seti olarak ikiye ayrılır.
2. Test seti, eğitim setiyle test edilir ve başarısı hesaplanır.(T1)
3. Eğitim setimizin üzerine ek eğitim setindeki veriler eklenerek yeni eğitim seti oluşturulur.
4. Test setimiz oluşturduğumuz yeni eğitim seti ile test edilir ve başarısı hesaplanır.(T2)
5. T2'nin T1'e göre başarısı yüzde olarak hesaplanır.
6. Belirlenen sayı kadar testler yapıldıktan sonra her bir test başarı yüzdesi toplanıp test sayısına bölünürse transfer sonucu ortalama başarı artışı bulunmuş olur.

Çalışmamızda veri setlerini analiz edebilmek Weka [4] kütüphanesinde bulunan SMO, Random Forest, J48, Ibk ve Naive Bayes sınıflandırıcıları kullanılmıştır. Her bir şirket için her bir arff dosyası 10'ar kez test edilmiştir, sonuçlar ölçülmüştür. Sonuçlar üzerinden metin sayısallaştırma tekniklerinin, sınıflandırıcıların başarıları gruplandırılmış ve Tablo 2, Tablo 3'te gösterilmiştir.

Tablo 2. Turkcell-Vodafone Veri Setinde Metin Sayısallaştırma Yöntemlerinin Ortalama Başarı Artışı

Metin Temsil Yöntemi	Tf Dosyalar Ortalama Artış	Binary Dosyalar Ortalama Artış
3 gram	4,18	2,95
Orijinal Kelime	1,8	7,55
Kelime Kökleri	1,82	2,53
Kelime Kökleri + Kelime Öneri	2,35	4,36

Tablo 2'ye göre Turkcell'den Vodafone'a ve Vodafone'dan Turkcell'e transfer testleri sonucu başarıyı en fazla artıran metin sınıflandırıcılar orijinal kelime ile binary tipinde oluşturulan ve kelime kök+ kelime öneri ile binary tipinde oluşturulan arff dosyaları olmuştur ve bu arff dosyaları 100'er test için seçilmiştir.

Tablo 3. Turkcell-Vodafone Veri Setinde Sınıflandırıcıların Ortalama Başarı Artış Oranları

Sınıflandırıcılar	Transfersiz Ortalama Başarı	Transfer Sonucu Ortalama Başarı Artışı %
Random Forest	80,08	2,61
SMO	83,14	1,51
J48	78,89	2,41
Naive Bayes	78,23	-0,51
Ibk	60,33	11,04

Oluşturulan Turkcell-Vodafone veri setini Weka içerisinde bulunan zeroR ile test yaptığımızda %69 sonucu elde edilir. Bu değer veri setinin en kötü şartlarda elde etmesi gereken başarı oranıdır. Tablo 3'e göre başarıyı en fazla artıran sınıflandırıcı Ibk'dır. Fakat Ibk ile yapılan testlerde sonuçlar genel olarak temel başarı oranını geçemediğinden, Ibk başarılı değildir denilebilir. Bu yüzden 100'er test için ilk olarak Random Forest seçilmiştir. Daha sonra gelen J48 sınıflandırıcımız da Random Forest'a çok benzediğinden SMO 100'er test için ikinci sınıflandırıcı olarak seçilmiştir. Metin sınıflandırıcıları seçtikten sonra elimizde orijinal kelimelerle binary tipinde, kelime kökleri ve kelime önerileriyle binary tipinde oluşturulmuş veri setleri kalmıştır. Bu veri setleri Turkcell'den Vodafone'a ve Vodafone'dan Turkcell'e transfer edilecek şekilde oluşturularak 4 veri seti elde edilmiştir. Bu veri setlerinin her biri SMO ve Random Forest ile transfer yöntemleri kullanılarak 100'er test yapılmıştır. Her bir veri setinin testleri

sonucunda, transfer yöntemlerinin o veri seti üzerindeki başarıları 1-15 arası numaralandırılmıştır. Daha sonra Turkcell-Ttnet veri tabanı orijinal kelimelerle ve binary yöntemiyle her iki şirket için de veri setleri oluşturulmuştur. Bu veri setlerinin de her birine de SMO ve Random Forest ile 100'er kez test yapılmıştır. Her bir veri seti testinin sonucunda transfer yöntemlerinin başarıları sıralanmıştır. En başarılı transfer yöntemini seçmek için transfer yöntemlerinin başarı sıralarının aritmetik ortalaması alınarak Tablo 4'te gösterilmiştir. Ayrıca transfer edilen ortalama tweet sayıları da Tablo 4'te belirtilmiştir.

Tablo 4 Transfer yöntemleri başarı sırası

Transfer Yöntemi	Turkcell-Vodafone Ort. Başarı Sırası	Ort. Trnsfr Sayısı	Turkcell-Ttnet Ort. Başarı Sırası	Ort. Trnsfr Sayısı
Nötr Olmayan	4,75	1484	15	1375
Firma adı geçmeyen	7,75	1414	2,5	1957
Eş-zıt tweet ekleme	2,25	2132	4	2685
Hem nötr hem firma geçme	5	1401	14	1342
Bfs analiz kelime geçme>0	7,875	1352	9,5	1580
Bfs analiz kelime geçme<2	7,5	882	5,75	1559
Ki Kare analiz kelime>0	8,375	1060	9	1389
Ki Kare analiz kelime<2	7,125	844	6,5	1365
Kosinüs benzerlik en benzer %50	13,125	750	10,25	1000
Kosinüs benz en az benzer %50	10,625	750	6	1000
Kosinüs benz en benzer %25 az benzer %25	7,375	750	7,5	1000
Korelasyon benz en ilişkili %50	11,5	750	10,25	1000
Korelasyon benz en az ilişkili %50	12	750	7,25	1000
Korelasyon benz en ilişkili %25 en az ilişkili %25	9,625	750	7	1000
Hepsi	5,125	1500	5,5	2000

4. Sonuçlar

Duygu analizinin temel problemlerinden bir tanesi etiketli verilerin eksikliğidir. Verilerin doğru bir şekilde analiz edilebilmesi için yeterli miktarda etiketli veri gereklidir. Fakat etiketsiz verileri etiketlemek oldukça maliyetli bir iştir. Bu çalışmada aynı sektörde olan iki firma arasında bir firmanın sahip olduğu etiketli verilerin diğer firma için de kullanılabilmesini sağlamak için çalışmalar yapılmıştır. Bu çalışma kapsamında veri seti kelime kökleri, 3 gram, orijinal kelimeler ve kelime kökleri ile önerilen kelimeler metin sayısallaştırma yöntemlerini kullanarak veri setleri oluşturulmuştur. Her iki firma için de SMO, Random Forest, J48, Naive Bayes ve Ibk sınıflandırıcıları kullanılarak 10'ar test yapılmış ve orijinal kelime binary, kelime kökleri ve önerilen kelimeler binary arff dosyaları ile Random Forest ve SMO sınıflandırıcıları başarılı olarak seçilmiştir. Seçilen veri setleri her bir firma için seçilen sınıflandırıcı ve transfer yöntemleriyle 100'er kez test edilmiştir. Testler sonucunda transfer öğrenme konusunda en başarılı sınıflandırıcı Random Forest, en başarılı sayısallaştırma yöntemi ise orijinal kelimelerle oluşturulmuş binary veri seti olmuştur. Transfer öğrenme yöntemlerinde ise Turkcell-Vodafone veri seti için veri setimize ek eğitim seti haricinde eğitim setinden eş-zıt anlamlı tweetleri üretip eklediğimiz yöntem, Turkcell-Ttnet veri setinde ise Turkcell ve Ttnet isimlerinin aynı anda geçmediği verilerin transfer edildiği yöntem en başarılı olmuştur. Genelde ek eğitim setindeki tweetleri transfer ettiğimizde başarının arttığı görülmüştür.

5. Kaynaklar

[1] Raina, R.,Koller, D. et. Al., (2006) "Constructing Informative Priors using Transfer Learning", 23rd International Conference on Machine Learning, Pittsburgh.

[2] Çetin, M., Amasyalı M.F., (2012) "Active Learning for Turkish Sentiment Analysis", Yıldız Technical University Department of Computer Engineering.

[3] "Türkçe Anlamsal İlişkiler Veri Kümesi", http://www.kemik.yildiz.edu.tr/data/File/TDK_viki.rar , 2016.

[4] Hall, M., Frank, E. et. Al. (2009) "The Weka Data Mining Software in Java", http://www.cs.waikato.ac.nz/ml/weka/".

[5] "Turkish Natural Language Processing with Zemberek",http://www.java2s.com/Code/Jar/z/Downloadzemberekcekirdek21jar.htm, 2016.