Spor Metinleri Sınıflandırma

Bülent Siyah

MKÜ Bilgisayar Mühendisligi Bölümü
bulentsyh@gmail.com

Özetçe

Bu çalışmada, Türkçe'nin yapısı kullanılarak türü bilinmeyen Türkçe bir metnin sınıflandırılması için geliştirilmiş bir özellik çıkarma yöntemi anlatılmaktadır. Metne ait olan özellik değerleri, o doküman içerisinde geçen kelime gövdelerinin her sınıf içerisindeki kullanım sıklılarının toplamından oluşmaktadır. Kelime gövdesi hangi sınıfta daha çok geçmiş ise ilgili metin bu sınıfa daha fazla dahil olmaktadır [1].

Bu çalışmada futbol, basketbol, tenis olmak üzere üç sınıf seçilmiştir. Elde edilen özellik vektörünün başarısı, Naive Bayes sınıflandırma yöntemi ile en yüksek %80 olarak alınmıştır.

1. Giriş

Günümüzdeki gelişmeler veri birikiminin artmasına neden olmaktadır. Bu artışla istenilen verilere ulaşabilmek için metinlerin sınıflandırılması ihtiyacı doğurmuştur. Metin sınıflandırma, o metnin özelliklerine bakarak önceden belirlenmiş belli sayıda kategorilerden hangisine dahil olacağını belirlemektir. Metin sınıflandırma bilgi alma, bilgi çıkarma, döküman filtreleme, otomatik olarak metadata elde etme ve web sayfalarını hiyerarşik olarak düzenleme gibi pek çok alanda önemli rol oynamaktadır [2]. İnternet'in ilk yıllarında kategorilendirme işlemi arama motorlarındaki uzman toplulukları tarafından elle gerçekleştirilmekteydi. Ancak internet'teki bliginin bugün ulaştığı boyut ve daha belkide daha önemlisi bilginin artış miktarı bu kategorilendirme(sınıflandırma) işleminin elle yapılmasını imkansız hale getirmiştir. İşte bu problemi çözebilmek için otomatik metin sınıflandırma sistemleri ortaya çıkmaya başlamıştır. Bu çalışmada içerisinde üç farklı kategoride 60 adet dökümandan oluşan bir veri seti kullanılmıştır.Makine öğrenmesi yöntemlerinden Naive Bayes kullanılarak bir metnin türünün belirlenmesi gerçekleştirilmiştir.

2. Yöntem

Bu çalışmada spor metinlerin WEKA paket programı ile Naive Bayes sınıflandırma algoritmasında kullanılabilinmesi için arff oluşturucu program (www.atasoyweb.net/blog/programlarim-k4s0/metin-siniflandirma-icin-arff-olusturucu-y123.html) yardımı ile her kelimenin her sınıfta kaç kez geçtiği ve her sınıf içerinde kaç farklı dökümanda bulunduğu, bu kelimelerin oranları ile öznitelikleriniz çıkarılması yapılmıştır.

2.1. Veri Seti

Sistemin eğitim ve test aşamalarında kullanılan veri seti HaberTürkSpor (www.htspor.com), (www.sporyazarlari.com) TenisHaber (www.tenishaber.com), BasketbolHaber (www.basketbolhaber.com), gibi sitelerin spor konularında yazı yazarlarının köşe yazılarından oluşturulmuştur. Eğitim setlerinde, her sınıftan 20 adet, 10 adet, 5 adet doküman alınmak şartıyla 60 , 30 ve 15 adetlik doküman setleri oluşturulmuş, test setinde ise her sınıftan toplamda 20 adet doküman sayısından geriye kalanlar kullanılmıştır. Böylece 30 ve 45 lik test setleri oluşturulmuştur. Toplamda 60 olan veri setleri 19239 kelimeden oluşmaktadır. Bu kelimelerden 900 tanesi öznitelik olarak alınmıştır.

*Tablo 1:*Spor Metinleri Sınıfları ve Eğitim ile Test Verileri Sayısı

Spor Kategorileri	Tenis	Futbol	Basketbol	Toplam
1.Veri Grubu **Eğitim/Test**	**10/10**	**10/10**	**10/10**	**30/30**
2.Veri Grubu **Eğitim/Test**	**5/15**	**5/15**	**5/15**	**15/45**
3.Veri Grubu Eğitim	**20**	**20**	**20**	**60**

Bu veri setleri ile 3 farklı veri grubu ile ilgili sınıflandırmalar yapılacaktır. Bunun dışında WEKA paketi ile öznitelikleri seçilmiş yani azaltılmış (900 olan öznitelik sayısı 27 indirgenmiş) 4.Veri grubu da veri setleri içinde bulunmaktadır.

2.2. Naive Bayes

Naive Bayes algoritması sınıflandırıcı bir algoritmadır. Metin dökümanlarının sınıflandırılmasında yaygın olarak kullanılır. Uygulanabilirliği ve performansı ile ön plana çıkan bir algoritmadır. İstatistiksel yöntemler yardımı ile sınıflandırma yapar.

Naive Bayes algoritmasının uygulanmasında bir takım kabuller yapılır. Bunlardan en önemlisi niteliklerin birbirinden bağımsız olduğudur. Eğer nitelikler birbirini etkiliyorsa burada olasılık hesaplamak zordur. Niteliklerin hepsinin aynı derecede önemli olduğu kabul edilir.

Naive Bayes algoritması bit ağırlıklandırma yöntemi ile ve frekans ağırlıklandırma yöntemi ile kullanılabilir.

2.2.1. Naive Bayes algoritmasının bit ağırlıklandırma ile kullanımı

$$p(d|c_j) = \prod_{t=1}^{|V|} p(w_t|c_j)^{x_t}(1-p(w_t|c_j))^{(1-x_t)}$$

$$p(w_t|c_j) = \frac{1+B_{jt}}{2+|c_j|}$$

Şekil 1: Bit ağırlıklandırma formülü

Yukarıdaki denklemler ile d vektörünün cj kategorisinde olma olasılığı hesaplanır.
|V|: Sözlükteki kelime sayısı
Bjt: cj kategorisinde bulunan ve wt kelimesini içeren eğitim dokümanı sayısı
|Cj|: cj sınıfında bulunan eğitim dokümanı sayısı
Xt: Kelimenin ağırlığı(1veya0)

2.2.2. Naive Bayes algoritmasının frekans ağırlıklandırma ile kullanımı

$$p(d|c_j) = p(|d|)|d|!\prod_{t=1}^{|V|}\frac{p(w_t|c_j)^{xt}}{x_t!}$$

$$p(w_t|c_j) = \frac{1+N_{jt}}{|V|+N_j}$$

Şekil 2: Frekans ağırlıklandırma formülü

d: Kategori Sayısı
Njt: j sınıfındaki dokümanlar için de t kelimesinin görülme sıklığı
Nj: j sınıfındaki toplam kelime sayısı
P(|d|): Kategori olasılığı
Xt: Kelimenin frekansı
|V|: Kelime sayısı

Daha sonra bir önceki hesaplamada olduğu gibi M(C) değerleri hesaplanır. Doküman, M(C) değeri en büyük olan kategoriye ait olarak belirlenir [3].

2.3. Uygun Algoritma Seçimi

Algoritma ilk önce belirli bir training(eğitim) data'sı ile eğitilir. Daha sonra eğitilen bu algoritma(Genellikle test datasının yarısı hacminde bir data ile) testing(test) datası ile test edilir. Buradaki amaç daha önceden eğitilmiş algoritmanın test datasında bulunan kriterler ile test datasında bulunan sonuçları doğru tahmin edip edememesinin incelenmesidir. Buda algoritmanın başarısı anlamına gelmektedir.

Bu kısımda algoritma tarafından üretilen değerlerin TruePositive, TrueNegative, FalseNegative, FalsePositive sayılarına göre değerlendirilmesi gerekmektedir.
TP : Olumlu sonucu olan ve olumlu öngörülmüş örnek sayısıdır.
FN: Olumlu sonucu olan ve olumsuz öngörülmüş örnek sayısıdır.
FP : Olumsuz sonucu olan ve olumlu öngörülmüş örnek sayısıdır.
TN: Olumsuz sonucu olan ve olumsuz öngörülmüş örnek sayısıdır.
Bu oranlar ile ilgili örnek bir tablo aşağıdadır.

Tablo 2: Doğruluk ve Tahmin tablosu

	Tahmin		
Doğru	Sonuç	1	0
	1	TP	FN
	0	FP	TN

Şimdi ise doğruluk ve hata oranı hesaplamalarına bakalım.
Doğruluk = (TP+TN)/(TP+FN+FP+TN)
Hata Oranı = (FN + FP)/(TP+FN+FP+TN) [4].
Naive Bayes algoritmasının formülü:

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\,p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}.$$

3. Sonuçlar

Sistemin eğitilmesi için Türkçe gazetelerin web sitelerindeki 3 farklı konudan 20'şer adet spor metni alınmıştır. Bu metinler arff oluşturma programı yardımıyla dönüştürülmüştür. Sınıflandırma algoritmaları için WEKA paketi kullanılmıştır.

1.Veri seti Naive Bayes ile sınıflandırıldığında; 1.Veri seti eğitim için 30 adet, test içinse eğitim setinde olmayan 30 adet makale bulunan veri setidir.

Tablo 3: 1.Veri Setinin Değerlendirmesi (Naive Bayes)

Correctly Classified Instances	24	% 80
Incorrectly Classified Instances	6	% 20

Tablo 4: 1.Veri Setinin Hata Matrisi (Naive Bayes)

Gerçek Tür	**Tahmin Edilen Tür**		
	Tenis	Futbol	Basketbol
Tenis	10	0	0
Futbol	0	9	1
Basketbol	2	3	5

Toplam 30 test spor metninin 24 tanesini doğru bir şekilde sınıflandırılmıştır. Burada tüm hata oranı %20 olmuştur. En yüksek sınıflandırma oranı %100'lük başarı ile tenis içerikli spor metinleri olmuştur. Tenisten sonraki en doğru sınıflandırma futbol makaleleri olmuştur. 10 Adet futbol metinlerinden sadece 1 tanesi basketbol kategorisinde zannedilmiştir. Basketbol metinlerinde yarı yarıya doğruluk elde edilmiştir. Bu metinlerden 3 tanesi futbol, 2 tanesi de tenis zannedilmiştir.

2.Veri seti Naive Bayes ile sınıflandırıldığında; 2.Veri seti eğitim için 15 adet, test içinse eğitim setinde olmayan 45 adet makale bulunan veri setidir.

Tablo 5: 2.Veri Setinin Değerlendirmesi (Naive Bayes)

Correctly Classified Instances	36	% 80
Incorrectly Classified Instances	9	% 20

Tablo 6: 2.Veri Setinin Hata Matrisi (Naive Bayes)

Gerçek Tür	**Tahmin Edilen Tür**		
	Tenis	Futbol	Basketbol
Tenis	15	0	0
Futbol	0	11	4
Basketbol	0	5	10

Toplam 45 test spor metninin 36 tanesini doğru bir şekilde sınıflandırılmıştır. Burada tüm hata oranı %20 olmuştur. En

yüksek sınıflandırma oranı %100'lük başarı ile yine tenis içerikli spor metinleri olmuştur. Daha öncede gözlendiği gibi futbol içerikli metinler basketbol içerikli metinlere kıyasla daha doğru sınıflandırmışlar.

3.Veri seti Naive Bayes ile sınıflandırıldığında; 3.Veri seti eldeki tüm metinlerin eğitim için oluşturulan veri setidir.

Tablo 7: 3.Veri Setinin Değerlendirmesi (Naive Bayes)

Correctly Classified Instances	30	% 100
Incorrectly Classified Instances	0	% 0

Tablo 8: 3.Veri Setinin Hata Matrisi (Naive Bayes)

Gerçek Tür	**Tahmin Edilen Tür**		
	Tenis	Futbol	Basketbol
Tenis	10	0	0
Futbol	0	10	0
Basketbol	0	0	10

Şimdiye kadar olan sonuçlarda hep test ve eğitim verisi ayrı ayrı verilerek sonuçlar gözlemlendi. Eldeki toplam 60 spor metni ile eğitip yine bu eğitim verilerinden toplam 30 tanesini test için ayırdığımda sonuç %100 olmuştur. Bu sonuçta gözlemlediğim kendi eğitilen veride kesinlikle hatalı sınıflandırma yapmıyor olmasıdır.

4.Veri seti Naive Bayes ile sınıflandırıldığında; 4.Veri seti 900 olan öznitelik sayısını WEKA (selected attributes) ile 27 düşürüp tekrar oluşturulan veri setidir.

Tablo 9: 4.Veri Setinin Değerlendirmesi (Naive Bayes)

Correctly Classified Instances	30	% 100
Incorrectly Classified Instances	0	% 0

Tablo 10: 4.Veri Setinin Hata Matrisi (Naive Bayes)

Gerçek Tür	**Tahmin Edilen Tür**		
	Tenis	Futbol	Basketbol
Tenis	10	0	0
Futbol	0	10	0
Basketbol	0	0	10

Spor metinlerinden çıkarılan öznitelik sayısı 900 olmuştur. WEKA paket programın bir özelliği olan öznitelik seçme işlemi ile aslında bu işin daha az öznitelikle yine aynı sonuca ulaşacağını anlaşılmıştır. WEKA selected attributes sekmesi ile veriler taranınca 900 olan öznitelik sayısını 27 tane öznitelik ile tekrar arff dosyamı oluşturdum. Oluşturuduğum arff dosyası ile az önce yaptığım eğitim verilerinde kullanılan 30 tane test verisini aldığımda sonucun yine %100 olduğunu gözlemledim. Böylece eğitim ve test verilerinde kullanılan veri kaynakları aynı metinler ise çok daha az öznitelik ile aynı sonuçların ortaya çıktığı anlaşılmış oldu.

2.Veri seti Naive Bayes Multinomial ile sınıflandırma yapıldığında başarı oranı %77.7 , Naive Bayes Multinomial Updateable ile %77.7 , Naive Bayes Updateable ile %80 oranlarında başarılı sınıflandırılmışlardır. Değişik eğitim ve test setleri ile en yüksek başarı Naive Bayes ile %80 olmuştur. Diğer sınıflandırma türleri arasında da en yüksek başarı en çok kullanılan sınıflandırma metodu olan Naive Bayes ile elde edilmiştir.

4. Kaynakça

[1] H.Kemal Yıldız, Murat Gençtav, Nurullah Usta, Banu Diri, M.Fatih Amasyalı "Metin Sınıflandırmada Yeni Özellik Çıkarımı" 2009.

[2] Türkoğlu F., Diri B., Amasyalı F., "Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi", Turkish Symposium on Artificial Intelligence and Neural Networks, 2006.

[3] M.Ali Demir, "Naive Bayes Sınıflandırma Algoritması", 2010.

[4] Tamer ÖZ ," NaiveBayes Kullanarak DataMining I", 2007.