TEKNOLOJİ HABERLERİNİN EMOJİLERE(TEPKİLERE) GÖRE NAİVE BAYES SINIFLANDIRMA VE C4.5 KARAR AĞACI İLE HABER KATEGORİSİ ANALİZİ

Ali ZORLU

ÖZET

Günümüz internet ve sosyal medya teknolojisinde jest ve mimiklerimizin yerini artık emojiler aldı. Çoğu zaman, emojiler kişilerin mimik ve tepkileri yerine kullanılsa da gelişen teknoloji sayesinde bu emojiler; kişilik, duygu analizi yerine kullanılmaktadır.

Bu çalışmada ise teknoloji haberlerdeki emojilerin (sinirli, kızgın, üzgün, şaşırma, sevinme) haber kategorisi üzerindeki etkisini sınıflandırma ve karar ağacı algoritmaları ile analiz edilmiştir.

Son olarak genel bir değerlendirme ve sonuç yazılarak çalışma sonlandırılmıştır.

Anahtar Kelimeler: *Emoji, Bayes, C4.5 Karar Ağacı, Haber, Kategori Analizi*

ABSTRACT

In today's internet and social media technology, gestures and mimics are now replaced by emojis. Often, though emojis are used instead of mimic and reaction of people, thanks to the developing technology these emojis; personality, emotion analysis.

In this study, the emojis in the news (nervous, angry, sad, surprised, rejoicing) were analyzed by classification of the effect on the news category and decision tree algorithms.

Finally, the study was terminated by writing a general evaluation and conclusion

Keywords: *Emoji, Bayes, C4.5 Decision Tree, News, Category Analysis*

1. GİRİŞ

Sembolik iletişim dili haline gelen emoji, Japonca kökenli bir kelime olarak hayatımıza yerleşmiş olup, “E” resim, imge ; “Moji” ise karakter, kişilik, özne anlamını taşımaktadır.

Emoji, Japon Mobil Operatörü NTT DoCoMo şirketi yazılımcısı Shiegetaka Kurita, tarafından özgün ve iletişime katkı amaçlı yapılan bir çalışma ile gerçek hayattaki insanların yüz ifadelerinden esinlenerek 15X15 boyutunda duygu içeren emojiler üretmesiyle başladı.[1]

Günümüzde ise birçok teknolojik pazara sahip şirketler kendi emojilerini üreterek durum, resim, yorum gibi içerik ve anlam yoğunluğu büyük uygulamalarda kullanıma sokmuş halde.

Emojinin teknolojideki yeri ise kullanıcıların, ziyaretçilerin, halkın duygu, kişilik, karakter zevk ve beğenileri gibi birçok analizi yapacak duruma kadar gelmiştir ve kullanılır haldedir. Birçok büyük sistemler bu teknolojiyi politik pazarlamadan tutun veri pazarı haline kadar getirmiş ve kullanıcıların ücretsiz kullanımına sunulan sistemlerden ticari pazar hacmi üretmişlerdir.

Bu makalede ise haberlerin kategorileri(toplam 18 kategori) analizini verilen emoji tepkilerinin sayılarına göre(bunlar şu şekildedir: iconShy(utanma), iconAngry(kızgın), iconLaugh(gülme), iconSad(üzgün) , iconAmazing(şaşırma)) analizi naive bayes sınıflandırma ve c4.5 karar ağacı algoritmaları kullanılarak yapılmıştır.

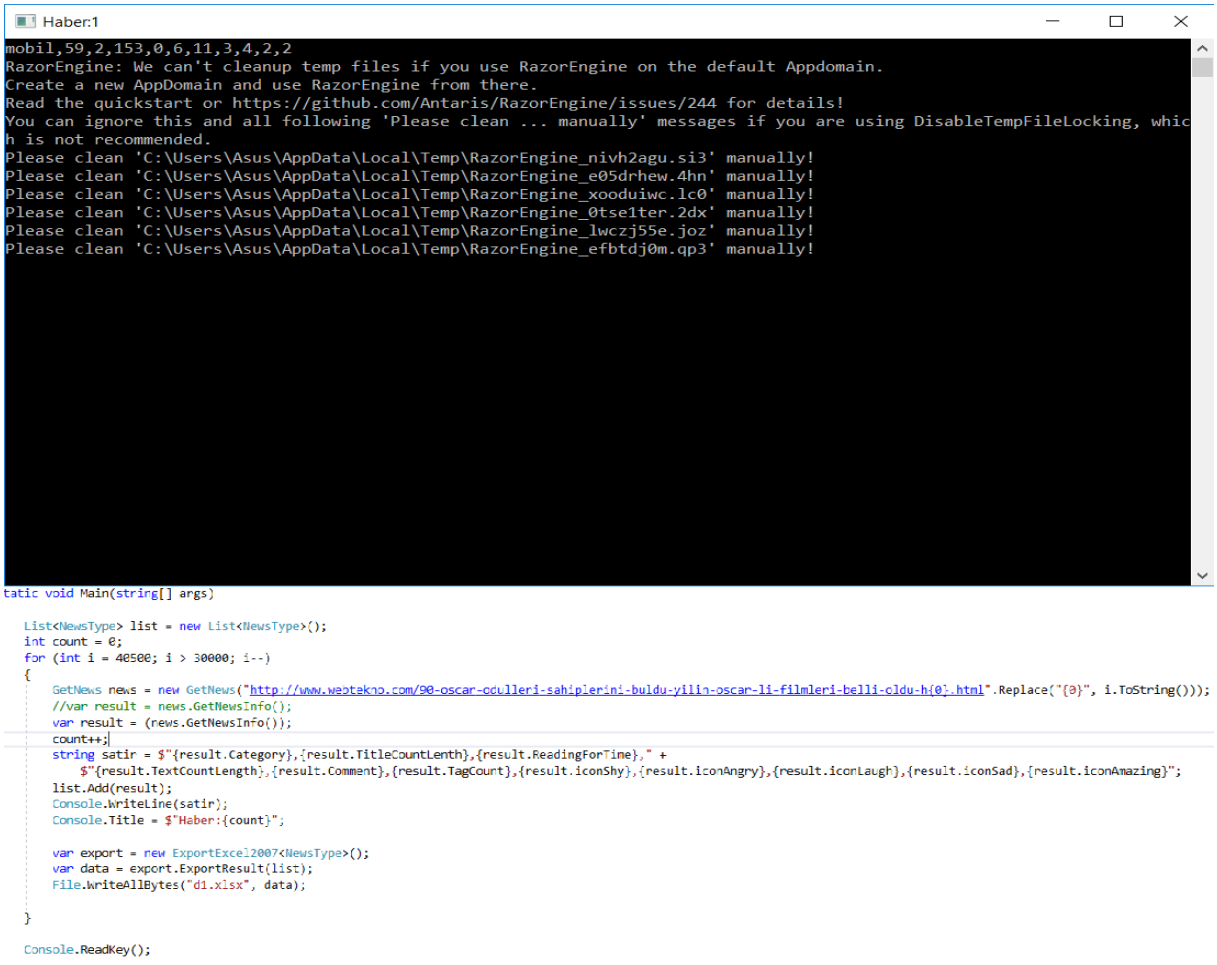
2. ÖRNEK VERİ SETİ VE VERİ SETİNİN ELDE EDİLME SÜRECİ

Bu çalışmada kullanılan veriler webtekno.com adlı teknoloji haber sitesinden bir program yazılarak veriler alınmış ve emoji tepkileri sayısal halde getirilmiştir.

Program, .Net dili ile HtmlAgilityPack, kütüphanesi aracılığı ile yazılmıştır. Programa ait kod ve çalışır örnek görüntü aşağıda verilmiştir.

NOT: Programın yazımı sırasında;

Haberin okuma süresi(kelime başına, ortalama dk.),haber etiketlerinin(tag) sayısı, haber yorum sayısı gibi diğer etmenler de hesaba katılmıştır ancak çalışmanın amacını farklı bir yöne sürüklediğinden ve bazı veriler sıfır(0) olarak geldiğinden, bunlar weka yazılımında analiz edilmeden önce kaldırılmıştır.



Toplam 4000 haber ile program sonlandırılmıştır. Ancak bazı veriler hatalı olduğundan; düzeltilerek hazır hale getirilmiş ve 3869 adet haber ile veri seti oluşturulmuştur.

Tablo 1'de 6 özellik içeren 14 veri içeren örnek bir set verilmiştir. Bu özellikler Category(Kategori), iconShy(utanma), iconAngry(kızgın),iconLaugh(gülme), iconSad(üzgün), iconAmazing(şaşırma) anlamını taşımaktadır. *Tablo 2*'de ise kullanılan veri setinde kategorilerin toplam örnek sayısı yer almaktadadır.

Tablo 1:14 veri içeren örnek veri seti

Category	iconShy	iconAngry	iconLaugh	iconSad	iconAmazing

Bilim-haberleri	13	2	1	1	4
Uzay	28	5	50	5	37
Mobil	4	34	1	1	1
Yasam	1	3	14	1	3
Sosyal-medya	9	21	4	3	3
İnternet	50	13	3	4	6
Sektorel	19	20	6	1	5
Kripto-para	3	4	35	2	12
Oyun	12	212	82	54	13
Zimbirtilar	6	2	3	1	43
Sosyal-medya	73	45	12	6	3
Giyilebilirteknoloji	1	18	1	0	3
Yazilim	14	0	0	1	3
Donanim	7	2	27	1	2

Tablo 2:Kullanılan veri setindeki toplam örnek sayısı

Category	**Örnek**
Mobil	806
Bilim-haberleri	260
Uzay	156
Yasam	386
Sosyal-medya	235
İnternet	399
Sektorel	328
Donanim	181
Otomobil	154
Kripto-para	157
Mobil-uygulama	85
Oyun	253
Zimbirtilar	57
Sinema	166
Yazilim	154
Yapay-zeka	43
Giyilebilirteknoloji	40
Kampanya	9
TOPLAM	3869

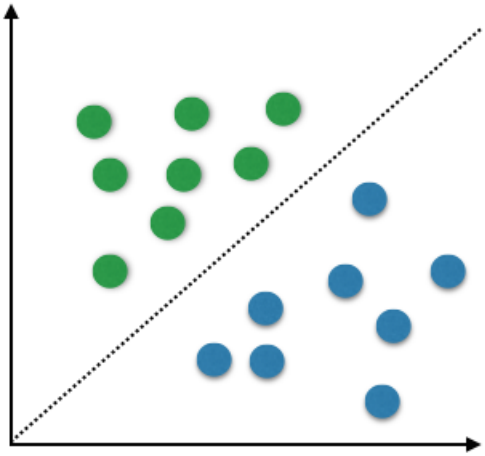
3. ÇALIŞMADA KULLANILAN METOTLAR

A-)NAİVE BAYES SINIFLANDIRMA ALGORİTMASI

Veri madenciliğinde, olasılık temelli sınıflandırma (classification) algoritması olarak bilinen Naive bayes, verilen özellikler arasındaki ilişkiyi tahmin eder.(OLGUN 2014:306)

Naive bayes algoritmasında öğrenme mantığı esastır. Verilen eğitim seti üzerinde özelliklerin sonuçlarını her defasında hesaplar. [2]

Bayes sınıflandırıcı ailesinde bulunan, naive bayes doğrusal(lineer) ve verimli bir sınıflandırıcı olarak bilinir. Saf(naive) adı da yine verilen veri seti üzerindeki özellikleri birbirinden bağımsız olarak hesaplamasından gelmektedir. Teoride bağımsızlık kavramı böyle iken; uygulama da ise bağımsızlık özelliği çoğunlukla ihlal edilir.[3]



Lineer (doğrusal) bir problemde sınıflandırma grafiği

Naive bayes algoritması öğrenme tabanlı bir algoritma olduğu için makine öğrenmesinde(machine learning),metin içerikli sınıflandırmalarda, mail spam filtresi oluşturulmasında kullanılır.

Naive bayes sınıflandırma algoritması, adını aldığı Thomas Bayes(1701-1761) tarafından formüle edilen basit bir olasılık modelleme üzerine kuruldu[4].Bu olasılık modeli şu şekilde verilebilir:

Özellik sayısı x olmak üzere

$$x=(x_1, x_1,\ldots, x_n)$$

Her bir k sınıfı için olasılık sonucumuz Ck olsun

$$P(Ck \mid x_1, x_2, \ldots, x_n)$$

Probleme ait koşullu olasılık şu şekilde elde edilebilir.

$$P(Ck \mid x) = \frac{P(Ck)P(x \mid Ck)}{P(x)}$$

Yukarıda verilen koşullu olasılık denklemi aşağıdaki gibi daha Türkçe halde yazılabilir.

$$Olasılık = \frac{Şartlı\ Olasılık\ .Önceki\ Olaslık}{Durum}$$

Özellik sayısı kadar yani x kadar koşullu olasılığın hesaplama işlemi ise şu şekilde yazılabilir.

$$P(\ x\ |Ci) = \prod_{k=1}^{n} P(\ xk\ |Ci)$$

Her sınıfa ait koşullu olasılık hesaplanırken de büyük olan sonuç işleme alınmalıdır bunu da şu şekilde formüle edilebilir.

$$\arg\max_{Ci}\{P(\ x\ |Ci)P(Ci)\}$$

B-)C4.5 KARAR AĞACI ALGORİTMASI

C4.5 algoritması entropiye (kümede ya da özelliklerdeki düzensizliğe) dayalı sınıflandırıcı algoritmalarından biridir.-Sınıflandırma algoritmalarının temelinde ise yine öğrenme mantığı yatmaktadır-.Karar ağaçları(decision tree) ile oluşturduğu kurallar sonucunda, kurallara uygun veriler sınıflandırılmış olur. [5]

Entropi, bir uzayda, olayda, kümede biçimsizliğin ölçüsü olarak bilinir.

Genel formülü ise

$$H = -\sum_{i=1}^{n} p_i \bullet \log_{2(p_i)}$$

Pi kümesindeki düzensizliğin sistem entropisi

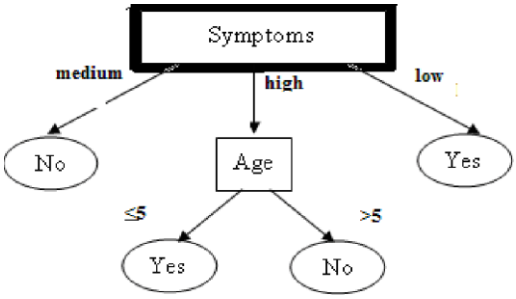
Veri seti üzerindeki bütün nitelikler için entropi hesabı yapılır ve sonunda ilgili niteliklerin kazanç ölçütleri(information gain) hesaplanır. Kazanç ölçütüne dair genel formül ise şu şekilde verilebilir;

$$IG(T,a) = H(T) - H(T|a)$$

IG(T, a):T niteliğinin kazanç ölçütü

Bulunan IG(T, a) veri setindeki bir özellik için kazanç ölçütünü niteler.

Bu kazanç ölçütleri veri setindeki bütün özellikler için hesaplanır ve ortaya çıkan kazanç ölçütleri içerisinde en büyük kazanç ölçütüne sahip(IG) nitelik ile karar ağacı oluşturulmaya başlanır.



Semptomlara dair örnek bir karar ağacı

4. DENEYSEL ÇIKTILAR

Bu çalışmada yer alan veri seti, WebTekno adlı haber sitesinden çekilmiş olup 18 haber kategorisi ve 5 farklı emoji tepkisi niteliklerini barındırmakta. Toplam 3869 veri içermekte.

Bu veri setine Naive Bayes algoritması uygulandığında;

3869 veriden %70 eğitim verisi olarak kalan %30'u da (1161) sınıflandırma için kullanıldı. Bu 1161 veri içerisinde düzenli sınıflandırma yapabildiği veri oranı %9.99(116 veri)

18 satır,18 sütunluk bir karışıklık matrisi(confusion matrix) oluştu. Matris görüntüsü aşağıdaki gibidir.

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r   <-- classified as
 40  7  0  5  0  9  9 14  0 37  0  0  2  1 44  0 90  0 |  a = mobil
  7 18  0  0  8  0  0  4  0  5  1  0  0  0  5  0 21  0 |  b = bilim-haberleri
  5  6  0  0  0  0  0  6  0  6  0  0  3  0  6  0 18  0 |  c = uzay
 12 27  0  0  7  6  3  7  0 11  2  0  4  1 10  0 19  0 |  d = yasam
  3  2  0  1  1  5  1  1  0 11  1  1  3  3 11  0 29  0 |  e = sosyal-medya
 13  8  0  2  3 11  1  4  0 19  4  0  4  1 16  0 43  0 |  f = internet
 10  3  0  0  2  1  7  4  0  8  1  0  0  1 18  0 40  0 |  g = sektorel
  3  2  0  0  0  2  1  6  0  3  1  0  0  0 10  0 22  0 |  h = donanim
  3  3  0  0  0  1  3  3  0  6  2  0  0  0  6  0 14  0 |  i = otomobil
  3  1  0  0  0  0  0  0  0 14  0  0  2  0  5  0 18  0 |  j = kripto-para
  2  0  0  0  0  0  0  1  0  4  0  0  0  1  6  0 15  0 |  k = mobil-uygulama
 11  2  0  0  1  4  0  9  0 14  0  0  5  2 15  0 16  0 |  l = oyun
  3  3  0  0  0  1  0  0  0  1  0  0  0  0  1  0  6  0 |  m = zimbirtilar
  3  4  0  0  0  2  3  5  0  9  0  0  3  1 12  0 18  0 |  n = sinema
  2  1  0  0  0  0  1  0  0  5  0  0  1  2 10  0 16  0 |  o = yazilim
  0  2  0  0  0  0  0  1  0  2  0  0  0  0  1  0  3  0 |  p = yapay-zeka
  0  0  0  0  0  0  0  1  0  1  0  0  0  0  2  0  8  0 |  q = giyilebilirteknoloji
  1  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0 |  r = kampanya
```

C4.5 algoritması uygulandığında ise düzenli sınıflandırma oranı % 16.1(187 veri) ya çıkmıştır. Ve karışıklık matrisi ise aşağıdaki gibi olmuştur.

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r   <-- classified as
 88 15 10 20 20 29 21  7  7  8  2 14  2  6  7  2  0  0 |  a = mobil
 11 10 11 12  1  7  5  2  0  2  0  3  0  1  4  0  0  0 |  b = bilim-haberleri
  9  7 11 10  1  2  3  3  1  0  0  3  0  0  0  0  0  0 |  c = uzay
 27  9  9 24  7 11  5  2  5  3  0  3  0  4  0  0  0  0 |  d = yasam
 21  5  4  5  6 11 13  0  0  5  0  0  0  1  2  0  0  0 |  e = sosyal-medya
 32  7  4 25  8 21 12  0  3  3  0  7  1  4  1  1  0  0 |  f = internet
 22  7  1  6  8 17 13  1  3  1  3  9  0  2  2  0  0  0 |  g = sektorel
 17  3  2  1  3  5  6  5  3  1  1  1  1  0  1  0  0  0 |  h = donanim
 16  2  1  8  2  6  4  0  0  1  0  1  0  0  0  0  0  0 |  i = otomobil
 16  1  1  5  5  4  4  1  0  2  0  2  0  0  2  0  0  0 |  j = kripto-para
 11  5  0  1  2  2  1  1  0  2  0  0  0  2  1  1  0  0 |  k = mobil-uygulama
 25  3  1 10  4 11  4  1  3  3  0  4  4  2  4  0  0  0 |  l = oyun
  5  2  0  1  0  1  2  0  1  1  1  0  0  1  0  0  0  0 |  m = zimbirtilar
 20  4  5  5  2  4  5  3  1  0  0  4  2  3  2  0  0  0 |  n = sinema
 17  2  0  2  1  4  5  1  2  0  0  1  0  3  0  0  0  0 |  o = yazilim
  3  3  0  1  0  1  0  0  0  0  0  1  0  0  0  0  0  0 |  p = yapay-zeka
  4  1  0  1  2  0  2  0  1  0  0  1  0  0  0  0  0  0 |  q = giyilebilirteknoloji
  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  0 |  r = kampanya
```

Sınıflandırma için ortaya konulan 1161 verinin 2 farklı yöntemdeki karşılaştırması *Tablo 3* 'te verilmiştir.

Category	**C4.5**	**Naive Bayes**
Mobil	**%34.1**	**%15.5**
Bilim-haberleri	**%14.5**	**%26.1**
Uzay	**%22.0**	**0**
Yasam	**%22.0**	**0**
Sosyal-medya	**%8.2**	**%1.4**
İnternet	**%16.3**	**%8.5**
Sektorel	**%13.7**	**%7.4**
Donanim	**%10.0**	**%12.0**
Otomobil	**0**	**0**
Kripto-para	**%4.7**	**%32.6**
Mobil-uygulama	**0**	**0**
Oyun	**%5.1**	**0**

Zimbirtilar	**0**	**0**
Sinema	**%5.0**	**%1.7**
Yazilim	**0**	**%26.3**
Yapay-zeka	**0**	**0**
Giyilebilirteknoloji	**0**	**%66.7**
Kampanya	**0**	**0**

Tablo 3:Haber kategori analizi için kullanılan sınıflandırma algoritmalarının TP(true positive)-Doğruluk yüzdelerinin karşılaştırılması

5. DEĞERLENDİRME VE SONUÇ

Emojilerle ziyaretçilerden olumlu-olumsuz feedback alan sistemler bunları bir çok parametrelerde kullanmakta olup, duygu analizi[7] kişilik analizi gibi bir çok veri madenciliği içeren yöntemler uygulanarak kullanılmakta. Bu çalışmadaki amaç da ziyaretçilerin emoji tepkileriyle ilgilendikleri kategorileri araştırmaktı.

Bu çalışmada emojiler vasıtası ile kategori sınıflandırması amaçlanmış ve 2 farklı sınıflandırıcı algoritma kullanılmıştır ve karar ağacına dayalı(entropi) algoritmanın olasılık tabanlı algoritmaya nazaran başarılı olduğu görülmüştür.

Bazı kategorilerde ise hiç sonuç alınamamıştır. Gözlem ve detaylandırma sonrasında ilgili kategoriler için yeterli haber bulunamamış ve sınıflandırıcıların bu kategorileri öğrenemediği görülmüştür. Çalışmanın son kısmında ise haber sayılarının kategori bazında eşit olmayan bir dağılımla hazırlandığı; haberlere tepki(emoji) veren kitlenin de hesaba katılması gerektiği göz önünden kaçırılmıştır.

6. KAYNAKÇA

1) ÖZÇEKİM, (2018.02.26). "Duygularımızın Tercümanı Emoji'nin Tarihçesi", Erişim tarihi: 2018.04.27, http://ozcekim.com.tr
2) OLGUN Mehmet, ÖZDEMİR Gültekin(2012,02),"İstatistiksel Özellik Temelli Bayes Sınıflandırıcı Kullanarak Grafiklerde Örüntü Tanımı",Journal of the Faculty of Engineering And Architecture Of Gazi University, 27, 2,303-311
3) Rish Irina , (2001), "An Empirical Study of the Naïve Bayes Classifier", IJCAI 2001 Work Empir Methods Artif Intell. 3,41-46.
4) WİKİ-ZERO, (2018.04.28). "Bayes Teoremi", Erişim tarihi: 2018.04.28,http://www.wiki-zero.net
5) Körting Thales, (2018). C4.5 Algorithm and Multivariate Decision Trees.
6) CAN Umit, ALATAS Bilal(2017,06),"Duygu Analizi ve Fikir Madenciliği Algoritmalarının İncelenmesi", International Journal Of Pure and Applied Sciences, 3, 1,75-111

DEĞİŞİKLİKLER

1. Yazım hatalarının ve noktalama işaretlerinin düzeltilmesi
2. Kaynakçadaki verilen tarihlerin düzeltilmesi
3. Sayfa biçimlendirmesi