

Film Yorumları ve Twitter Verileri Üzerinden Duygu Analizi

Halil İbrahim ÇELENLİ

Kocaeli Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Ana Bilim Dalı,
e-mail: 165112015@kocaeli.edu.tr

ÖZET

Günümüzde internetin hızlı bir şekilde gelişmesi ile verilerimiz sürekli olarak çoğalmaktadır. Bu durum teknolojinin gelişmesiyle birlikte verilerimiz içerisinde hızlı bir şekilde işlem yapabilmemizi amaçlamaktadır. Duygu analizi(Sentiment Analysis) hızlı bir şekilde insanların duyguları üzerinden çeşitli yöntemler açığa çıkarmaktadır. Söyleki : Bir kişinin 1000 adet film yorumu üzerinden 1001. Filmi beğenip beğenmeyeceği yorumunda bulunulabilmektedir. Bu sayede insanların beğeneceği filmler önerilebilir. Bu çalışma üzerinde 2 farklı model oluşturulup 2 farklı dataset üzerinden işlem yapılması sağlanmıştır. Duygu analizinde 2 seçenek verilmiş (olumlu , olumsuz) ve bunlar Naive Bayes yöntemi ile Python üstünden kodlamaları gerçekleştirilmiştir.

Anahtar Kelimeler:Naive Bayes; Duygu Analizi;Twitter Veri Analizi; Sosyal Medya

1. GİRİŞ

Teknoloji anlamında dünya hızlı bir şekilde ilerlemektedir. Bu ilerleme ulaşımı kolaylaştırdığı gibi verinin artmasına, insanların az sözle çok şey anlatmalarına da sebebiyet verebilmektedir. Sosyal Medya'da teknolojinin gelişmesi ile üzerinde durulması gereken önemli bir konu haline gelmiştir. Sosyal medya bir bilgi kaynağıdır ve bu bilginin analiz edilmesi gerekmektedir. Fakat veriler incelendiğinde çoğunluğun eksik hatalı yazıldığı görülmektedir. Bu nedenle veriler doğal dil işleme ile bir işlemden geçirilmesi gerekmektedir. Twitter, sosyal medya üzerinde veri analizi yapılabilecek en iyi ortamlardan birisi olarak nitelendirilebilir. Günümüzde twitter üzerinden toplanan veriler ile doğal dil işleme alanında çeşitli çalışmalar yapılmıştır. 2009 yılının Mayıs/Aralık ayı içerisinde

atılmış olan tweetlerden salgın hastalıkların önceden tahminleme (Yarrow & Sverdrup-Stueland, 2004) verilebilecek örneklerden birisidir. Bu işlemler yapılırken en iyi yol makine öğrenmesi(machine learning) yöntemlerini kullanmaktır.

Makine öğrenmesi geçmişten günümüze teknolojinin gelişmesine en büyük katkıyı sunan yapılardan birisidir. Bu yapı sayesinde bilgisayarlarda öğrenme yeteneğine sahip olup verileri kolay bir şekilde algılayıp sınıflandırabilmemizi sağlamışlardır.

Makine Öğrenmesinde 3 farklı öğrenme yöntemi bulunmaktadır. Bunlar gözetimli öğrenme(supervised learning), gözetimsiz öğrenme(unsupervised learning), pekiştirmeli öğrenme(semi-supervised learning) yöntemleridir. Çalışmamız da gözetimli öğrenmeyi(Supervised Learning) baz alarak bir çalışma gerçekleştirildi. Bayes teoremi gözetimli öğrenme yöntemlerinden birisi olarak nitelendirebiliriz. Duygu analizinde de kullanılan yöntemlerden birisidir.

Bayes teoremi duygu analizinin yanında çeşitli metin madenciliği(text mining) sınıflandırmalarında da kullanılır. Bunlardan bazıları : Yazar kimliği tanıma, yaş cinsiyet tanıma, spam filtreleme vs. verilebilir.

2. TEMEL BİLGİLER

2.1 Gözetimli Öğrenme

Gözetimli öğrenme, sürekli bir gözetimin olduğu makine öğrenmesi yöntemlerinden birisidir. Gözetimli öğrenme: Öğrenilmek istenen kavram ile toplanan gözlemlerin eğitim kümesi halinde öğreniciye verilmesi olarak da tanımlanabilir. İstenen her örnek için eğitim kümesinde çıktıya da yer verilir. Bu bilgiler sayesinde giriş ve çıkış arasında bir ilişki oluşur. İlişki sayesinde girdi kümesinden çıktı kümeleri tahmin edilebilir. Gözetimli öğrenmenin diğer öğrenme yöntemlerinden

farkı eğitim verisinin bulunmasıdır (Theodoridis & Koutroumbas, 2009).

2.2 Naive Bayes Sınıflandırıcısı

Temeli Bayes teoremine dayanmaktadır. Bir versiyonu olan Naif Bayes kullanılmaktadır ve bazı kabulleri vardır. Veri kümeleri arasında ilişki(biased) olmaması durumu beklenmektedir. Önemli özelliklerinden birisi de sisteme veri girişi oldukça kendini buna adapte edebilmesi ve eğitmesidir. Metin madenciliğinde Çeşitli yaklaşımları(Bag of words, N-gram vs.) kullanarak kolay bir şekilde sınıflandırma yapılabilir.

Temel Bayes kuralı Şekil-1 de gösterildiği gibidir.

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})}$$

$$p(\mathbf{x}) = \sum_{i=1}^2 p(\mathbf{x}|\omega_i)P(\omega_i)$$

Şekil 1: Bayes Kuralı

2 sınıflı bir bayes kuralında ise olasılık sonuçlarına göre verilen ifadenin hangi sınıfa ait olacağı Şekil-2 ile gösterilmiştir.

If $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$, \mathbf{x} is classified to ω_1
 If $P(\omega_1|\mathbf{x}) < P(\omega_2|\mathbf{x})$, \mathbf{x} is classified to ω_2

Şekil 2 : İkili Naive Bayes Sınıflandırması

Bayes kuralında temelde $p(\mathbf{x})$ olasılığını her sınıflandırma olasılığında aynı şekilde kullanacağımızdan, $p(\mathbf{x})$ ifadesi atılarak işlem yapılabilir.

Introduction to Information Retrieval (Manning, Raghavan, & Schütze, 2009) kitabında anlatıldığı üzere, c değerimiz sınıf ve d değerimizi de doküman olarak belirtirsek Maximum a Posteriori(MAP) değerimiz, bizim verimizi en iyi olasılık değeri veren sınıfamıza atamayı Şekil-3 üzerinde gösterildiği gibi sağlamaktadır.

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(c | d)$$

$$= \operatorname{argmax}_{c \in C} \frac{P(d | c)P(c)}{P(d)}$$

$$= \operatorname{argmax}_{c \in C} P(d | c)P(c)$$

Şekil 3 : MAP Hesabı

Naive bayes algoritmasının çalışma hızı diğer gözetimli öğrenme sınıflandırma algoritmalarına göre daha yavaştır. Bunun sebebi ise dinamik sistemlerin eğitilmesidir; daha sonra bu durumu kendi bünyesinde saklayamamaktadır. Çünkü her seferinde yeni verilerin olacağından olasılıkların değişmesi sebebiyle tekrar tekrar hesaplama ihtiyacının olmasıdır (Şeker, 2015).

Burada önemli olan noktalardan birisi de verilerimizin bağımsız olmasıdır. Verilerimiz bağımsız değilse yanlış sonuçlar alabiliriz.

2.3 Örnek Bayes Sınıflandırma

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Şekil 4 : Örnek Verilerimiz

Şekil-4 üzerinde görüldüğü gibi 4 adet eğitim ve 1 adet test kümemiz bulunmaktadır. Toplamda 2 adet sınıfamız bulunmaktadır. Burada Naive Bayes işlemini uygulayacak olursak :

$$\hat{P}(c) = \frac{N_c}{N} \quad \hat{P}(w | c) = \frac{\text{count}(w, c) + 1}{\text{count}(c) + |V|}$$

$P(c)$ = sınıfamızın etiketlediği belge/toplam belge

w = feature(öznelitikler)

c = sınıflarımız(c ve j)

Önsel $\Rightarrow P(c)=3/4$, $P(j)=1/4$

$P(\text{Chinese}/c) = (5+1)/(8+6) = 6/14 = 3/7$

$P(\text{Tokyo}/c) = (0+1)/(8+6) = 1/14$

$P(\text{Japan}/c) = (0+1)/(8+6) = 1/14$

$P(\text{Chinese}/j) = (1+1)/(3+6) = 2/9$

$$P(\text{Tokyo}/j) = (1+1)/(3+6)=2/9$$

$$P(\text{Japan}/j) = (1+1)/(3+6)=2/9$$

$$P(c/d5) \propto 3/4*(3/7)^3*1/14*1/14 \approx 0.0003$$

$$P(j/d5) \propto 1/4*(2/9)^3*2/9*2/9 \approx 0.0001$$

Sonuç olarak test verimizin c sınıfına ait olma olasılığı daha yüksek olduğu için 5 numaralı dokümanımız c sınıfına aittir diyebiliriz.

3.LİTERATÜR TARAMASI

1- Eyüp Sercan AKGÜL ve arkadaşları tarafından twitter verileri ile duygu analizi, n-gram ve sözlük modeli kullanılarak yapılmış. Gramlar içerisinde 3-gram iyi sonuç vermiştir. %72 doğruluk oranı ile sözlük modelinin gram modelinden daha iyi sonuç verdiği görülmüştür (Akgül, Ertano, & Diri, 2016).

2- Aysun GÜRAN ve arkadaşları 3 farklı dataset üzerinden Destek Vektör Makinelerini(DVM) kullanarak duygu analizi yapmışlar. En iyi doğruluk oranını %75 olarak belirlemişlerdir (Üran, Uysal, & Doğrusöz, 2014).

3-Cumali TÜRKMENÖĞLU tarafından n-gram'lar kullanılarak twitter verileri üzerinden %75 ile film yorumları veri kümesi üzerinden %79 oranında başarımlar sağlanmıştır (Türkmenoğlu, 2015).

4-Huma PARVEEN ve Shikka Pandey tarafından twitter verileri üzerinden Hadoop ile naive bayes algoritması kullanılarak duygu analizinde, verilerin ön işleme tabi tutulması ile iyi sonuçlar verdiği belirtilmiştir (Parveen & Pandey, 2016).

5-Shweta RANA ve Archana Singh tarafından twitter verileri üzerinden destek vektör makineleri ve naive bayes algoritmaları karşılaştırılmış. Destek Vektör Makinelerinin %75 ile daha iyi sonuç verdiği belirtilmiştir (Rana & Singh, 2016).

4.MİMARİ

4.1 Veri Analizi

Verilerimiz üzerinde gözetimli öğrenme işlemini gerçekleştireceğimiz için sistemimizi eğitmemiz için elimizde bir veri bulunması gereklidir. Gerekli olan veriyi 2 farklı şekilde ele almış bulunmaktayız. Film yorumlarında veri kümesi olarak Sentiment polarity dataset v2.0 üzerinden işlem yapılmıştır. Twitter veri kümesi olarak da NLTK kütüphanesi içerisinde json formatında pozitif ve negatif olarak ayrılmış veriler kullanılmıştır.

Sentiment polarity dataset v2.0 içerisinde pozitif ve negatif olarak ayrılmış 1000 adet pozitif ve 1000 adet negatif metin belgeleri bulunmaktadır. Şekil-5 üzerinde 1 adet örnek pozitif film yorumu gösterilmiştir.

Film adapted from comic books have had plenty of success, whether they're about superheroes (Batman, Superman, Spider-Man) or paired twisted kids (Sawyer) or the virtuous crowd for starters, it was created by Alan Moore and Eddie Campbell. Who brought the medium to a whole new level in the mid '80s with a 12-part series called the Watchmen. To say Moore and Campbell thoroughly researched the subject of Jack the Ripper would be like saying Michael Jackson is starting to look a little odd. The book (or graphic novel, if you will) is over 500 pages long and includes nearly 30 more that consist of nothing but footnotes. In other words, don't dismiss this film because of its source.

If you can get past the whole comic book thing, you might find another startling knock in from hell's directors, Albert and Allen Hughes. Getting the Hughes brothers to direct this seems almost as ludicrous as starting Sawyer up in, well, anything, but suffice me this: who better to direct a film that's set in the ghetto in question is, of course, Watchmen. In 1988 London's east end.

It's a filthy, smoky place where the streets are called "unfortunate" and are starting to get a little nervous about this mysterious psychopath who has been carving through their post when the first stuff turns up, copper Peter Guzik (Cuba Gooding Jr.) calls in Inspector Frederick Abnerline (Jimmy Smag, Alan) to crack the case. Abnerline, a widower, has prophetic dreams he unsuccessfully tries to quell with copious amounts of alcohol and spin.

Upon arriving in Watchmen, he befriends an informant named Nite II (another guinea, say it isn't so) and proceeds to investigate the horrifically gruesome crimes that even if he doesn't think anyone needs to be involved in Jack the Ripper, he's not going to let the particular here, other than to say Moore and Campbell have a unique and concerning theory about in the comic, they don't bother cloning the identity of the ripper, but someone's really happy (vertical line) and sad, yikes! (has this 7 cables) is a good job of deep it's funny to watch the locale blindly point the finger of blame at Jews and Indians because, after all, an Englishman could never be capable of committing such ghastly acts. And from hell's ending had me whistling the Stonecutters song from the Singapore for days (who holds back the electric car who made stonecutters a star?)

Don't worry - it'll all make sense when you see it.

Now onto from hell's appearance: it's certainly dark and kind enough, and it's surprising to see how much more it looks like a tin horizon film than planet of the apes did (at the time the genre is now used) completely finished (both color and music but not been finished), so no comment about watching scenes, but cinematographer Peter Deming (don't say a war scene unless Martin Scorsese) (Shakespeare in Love) production design turns the original proper surroundings into one creepy place.

Even the setting in from hell is solid, with the dreamy deep running in a typically strong performance and drolly handling a British accent. Jane Hula (you could's secret) and Richardson (100 skeletons) lay in great supporting roles, but the big surprise here is Graham. I cringed the first time she opened her mouth, imagining her attempt at an Irish accent, but it actually wasn't half bad.

The film, however, is all good.

Şekil 5 : Örnek Pozitif Film Yorumu

Twitter veri kümemiz ise pozitif ve negatif olarak 2 adettir. Veri formatları json tipindedir. Kendi twitter verilerinizi json formatına getirip onlar üstünden de veri analizi gerçekleştirebilirsiniz. Şekil-6 üzerinde json tipinde twitter üzerinden alınmış veriler gösterilmektedir.

```
{
  "coordinates": null,
  "coordinates": null,
  "text": "Apples for tea :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile background image url": "http://i
  },
  "coordinates": null,
  "coordinates": null,
  "text": "Everything in the kids section of IKEA is so cute. Shame I'm nearly 15 in 1 month :)",
  "user": {
    "screen name": "Bewildered",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "Beyonce that heart sliding into the waste basket :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "I still love Beyonce: I hate Japanese call him 'Vani' :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "They starting next week I love 'Vani' :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "Oh god, my babies' faces :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "Beyonce through male me mile :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "Beyonce Beyonce work neighbour on noters. Jaded why and he said hates the updates on search :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "Any? (I'm sorry: stalin) https://t.co/8j8j8j8j8j",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "Beyonce glacier was there in 1986 :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "I have a really good memory this but I'm never going to meet them :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "BeyonceBeyonce was there :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "BeyonceBeyonce happy trip, long solo, see you soon :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "It's so hard to believe :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "BeyonceBeyonce with love compliments they get you up early about the same day :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "Relate to the 'Power of Love' kind of 'We're all in your life... on my life... is full of them :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "Beyonce's please :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "BeyonceBeyonce in our own the :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "I feel stupid! just can't seem to grasp the basics of digital painting and nothing I've been researching is helping any :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "Good Lord :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "I feel lonely someone talk to me guys and girls :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "No assignment, but we have Project :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  },
  "coordinates": null,
  "coordinates": null,
  "text": "just want to play video games/watch movies with someone :)",
  "user": {
    "screen name": "pawpaw",
    "time zone": "Alaska",
    "profile": {
  }
}
```

Şekil 6 : Jsn Formatında Örnek Negatif Verileri

Verilerimizi analiz etmeden önce hangi

programlama dilinin problemimize uygun çözüm oluşturacağı düşünülmüş olup, tercihen python dili seçilmiştir. Python dilinin 2 farklı sürümü bulunmaktadır. Bunlar 2.x.x ile 3.x.x sürümleridir. Problemin çözümü python 3.x.x ile yapılmıştır. Python üzerinde analiz yapabilmeyi sağlayan araçları kapsayan Anaconda aracı tercih edilip, tool olarak jupyter tercih edilmiştir. Kütüphane olarak diğerlerine nazaran önem düzeyi yüksek olarak nitelendirilen Nltk(Natural Language Toolkit), kullanılmıştır. Nltk kütüphanesi Stanford Üniversitesi tarafından geliştirilen İngilizce metinlerde çok kullanılan bir araçtır (Bird, Klein, & Loper, 2009).

Duygu analizinde çeşitli sınıflandırmalar kullanılmaktadır. Yapılan literatür taramasında en iyi sınıflandırmalar KNN , Destek Vektör Makineleri ve Naive Bayes olarak belirlenmiştir. naive bayes algoritmamızı kullanabilmemiz için ilk olarak verilerimizi pozitif veya negatif olarak etiketlemeliyiz. Çözümümüzde Bag Of Words(kelime çantası modeli) seçildiği için her kelimemiz bizim birer vektör uzayımızda bulunan feature(öznitelik) olarak nitelendirilebilir. Vektör uzayımızı biraz azaltmak için veri kümelerimiz üzerinde yapılan işlemde Stop Words(etkisiz kelimeler) çıkarılarak işlem yapılmıştır. Ayrıca bazı özel karakterler de çıkartılarak işlem yapıldığında doğruluk oranı normal durum ile karşılaştırılmıştır.

Verilerimiz üzerinde bayesin bir modeli olan naive bayes üzerinden işlem yapılmıştır.

$$\frac{P(\text{etiket/features})}{P(\text{etiket}) * P(\text{features/etiket})} = \frac{P(\text{features})}{P(\text{features})}$$

P(etiket/features) => Posteriori(Sonsal)
P(etiket) => prior(önsel)
P(features/etiket) => Likelihood(olabilirlik)

$$\frac{P(\text{features/etiket})}{P(f1/etiket) * \dots * P(fn/etiket)} =$$

EtiketMAP(Etiketin Maksimum Sonsalı) = $\text{argmax} P(\text{features/etiket}) * p(\text{etiket})$

Yukarıda görülen formüllerde etiketlerimiz pozitif ve negatif olarak nitelendiririz. features

değerlerimizde belgelerimizde bulunan kelimelerimizdir.

4.2 Uygulama

Uygulama içerisinde 2 farklı veri kümesi üzerinde işlem yapıldığı için öncelikli olarak film yorumları üzerinden naive bayes algoritması kullanılıp daha sonrasında twitter verileri üzerinden naive bayes algoritması kullanılmıştır.

Film yorumları veri kümemiz üzerinden ilk olarak python ile verilerimiz çekilmiştir. Verilerimiz üzerinde etiketleme işlemi yapmadan önce stopwords kelimeler çıkartılıp sonrasında etiketleme işlemi yapılmıştır. NLTK kütüphanesinin naive bayes sınıflandırıcısını kullanabilmemiz için her kelimeyi etiketlememiz gerekmektedir. Kelimeler “True” olarak etiketliyoruz. Her belgemizi de Şekil-7 üzerinde görüldüğü gibi negatif ve pozitif olarak etiketliyoruz.

```
True, 'want': True, 'completely': True, 'final': True, 'five'
hrilling': True, 'engaging': True, 'meantime': True, 'really
flicks': True, 'actually': True, 'figured': True, 'half': Tr
, 'little': True, 'bit': True, 'sense': True, 'still': True,
ays': True, 'sure': True, 'given': True, 'password': True, '
ue, 'melissa': True, 'sagemiller': True, 'running': True, 'a
in': True, 'lazy': True, '!!': True, 'okay': True, 'people':
e, 'us': True, 'different': True, 'offering': True, 'insight
tor': True, 'chopped': True, 'shows': True, 'might': True, '
'turning': True, 'music': True, 'video': True, 'edge': True
'bentley': True, 'seemed': True, 'playing': True, 'exact': T
True, 'neighborhood': True, 'kudos': True, 'holds': True, 'e
ue, 'stick': True, 'entertain': True, 'confusing': True, 'ra
runtime': True, 'despite': True, 'ending': True, 'explanat
': True, 'slasher': True, 'flick': True, 'packaged': True, '
ds': True, 'also': True, 'wrapped': True, 'production': True
ever': True, 'whatever': True, 'skip': True, 'joblo': True
True, '/*': True, '10': True, 'blair': True, 'witch': True,
'stir': True, 'echoes': True, '8': True}, negative'
```

Şekil 7 : Negatif Olarak Etiketlenmiş Bir Veri

Verilerimizi pozitif ne negatif olarak etiketleyip aldıktan sonra Sınıflandırma işlemine geçeceğiz. Sınıflandırma işlemi için NLTK kütüphanesinin Naive Bayes Classifier metodunu kullanacağız. Toplam 1000 adet pozitif ve 1000 adet negatif veri kümemizin 1500 tanesini eğitim amaçlı kullanıyoruz. 500 tanesini de test amaçlı kullanıyoruz. İşlemlerimizi yaptıktan sonra Doğruluk(Accuracy) hesapladığımız zaman **%72.3** oranında bir doğruluk buluyoruz. Confusion(Karıştırma) Matrisine baktığımız zaman 250 adet negatif etiketli verinin 116 tanesini doğru 134 tanesini yanlış olarak sınıflandırmış. 250 adet pozitif etiketli verinin ise 246 tanesini doğru 4 tanesini yanlış olarak

sınıflandırmıştır. Şekil-8 üzerinden naive bayes sınıflandırmamızın sonuçları görülmektedir.

```

      | n | p |
      | e | o |
      | g | s |
      | a | i |
      | t | t |
      | i | i |
      | v | v |
      | e | e |
-----+-----
negative |<116>134 |
positive  | 4<246> |
-----+-----
(row = reference; col = test)

Accuracy :72.39999999999999

```

label	precision	recall	f_measure
negative	0.9666666666666667	0.464	0.627027027027027
positive	0.6473684210526316	0.984	0.780952380952381

Şekil 8 : Naive Bayes Sınıflandırması Sonucu

Stopwords kelimelerini çıkarmadan işlem yaptığımız zaman da **%71** oranında f1-score değerini hesaplıyoruz. Şekil-9 üzerinde gösterilmiştir.

Clasification:	precision	recall	f1-score	support
negative	0.96	0.48	0.64	250
positive	0.65	0.98	0.78	250
avg / total	0.81	0.73	0.71	500

```

Confussion matrix:
[[119 131]
 [ 5 245]]

```

Şekil 8 : Stop Words Kelimeler Çıkarılmadan Yapılan Naive Bayes Sınıflandırması

Şekil-10 üzerinde verilerimiz içerisinde modelimizin etkili bulduğu 20 adet özelliğimiz gösterilmiştir. Örnek olarak verilerimizde “magnificent” kelimesinin pozitif olarak etiketlenmesi negatif olarak etiketlenmesinden 15 kat daha fazladır.

Most Informative Features		
magnificent = True	positi : negati =	15.0 : 1.0
outstanding = True	positi : negati =	13.6 : 1.0
insulting = True	negati : positi =	13.0 : 1.0
vulnerable = True	positi : negati =	12.3 : 1.0
ludicrous = True	negati : positi =	11.8 : 1.0
avoids = True	positi : negati =	11.7 : 1.0
uninvolving = True	negati : positi =	11.7 : 1.0
fascination = True	positi : negati =	10.3 : 1.0
astounding = True	positi : negati =	10.3 : 1.0
idiotic = True	negati : positi =	9.8 : 1.0

Şekil 10 : Etkili 10 Özellik

Twitter verileri üstünde yaptığımız naive bayes sınıflandırmasında ilk olarak verilerimizi python üstünden çekeriz. Verilerimiz 2 adettir

bunlar json formatında bulunup, 5000 adet pozitif ve 5000 adet negatif olarak verilmiş verimiz bulunmaktadır. Twitter verileri 140 karakter ile sınırlı olduğu gibi Şekil-11 üzerinde göreceğiniz gibi kısa cümleler halindedir verilerimiz. Normal film yorumlarından bir farkı da verilerimiz içerisinde “smile” işaretlerinin bulunmasıdır. Bu işaretler de pozitif ve negatif olarak ayrılmış halde verilerimiz içerisinde bulunmaktadır. Şekil-11 üzerinde örnek olarak verilmiş 5 adet negatif olarak nitelendirilmiş twitter verisi bulunmaktadır.

```

hopeless for tmr :(
Everything in the kids section of IKEA is so cute. Shame I'm nearly 19 in 2 months :(
@Hegelbon That heart sliding into the waste basket. :(
"@ketchburning: I hate Japanese call him "bani" :( :("

```

```

Me too
Dang starting next week I have "work" :(

```

Şekil 11 : Negatif Olarak Verilmiş 5 Adet Veri

Twitter verilerimizi etiketleme işlemini gerçekleştirirken stopwords kelimeleri çıkartıyoruz. Kelimelerimizi çıkarttıktan sonra Şekil-12 üzerinde görüleceği gibi negatif ve pozitif olarak etiketliyoruz.

```

({'#FollowFriday': True, '@France_Inte': True, '@PKuchly57': True, 'rs': True, 'community': True, 'week': True, ':)': True}, {'positive'})

```

Şekil 12 : Pozitif Olarak Etiketlenmiş Bir Veri

Verilerimi etiketledikten sonra NLTK kütüphanesini kullanarak 7000 adet veriyi eğitim amaçlı kullanıp, 3000 adet veriyi test amaçlı kullanıyorum. Naive Bayes Classifier'ı uyguladığım zaman **%99.2** oranında bir doğruluk sağlıyorum. 1550 tane negatif verinin 1490 tanesi doğru 10 tanesi yanlış tahminde bulunmuştur modelimiz.1500 adet pozitif verinin ise 1486 tanesini doğru 14 tanesini yanlış tahmin etmiştir modelimiz.Şekil-13 üzerinden naive bayes sınıflandırmamızın sonuçları görülmektedir.


```

      | n | p |
      | e | o |
      | g | s |
      | a | i |
      | t | t |
      | i | i |
      | v | v |
      | e | e |
-----+-----+
negative |<1490> 10 |
positive | 14<1486> |
-----+-----+
(row = reference; col = test)

Accuracy:99.2

```

label	precision	recall	f_measure
negative	0.9906914893617021	0.9933333333333333	0.9920106524633822
positive	0.9933155080213903	0.9906666666666667	0.9919893190921228

Şekil 13 : Twitter Dataseti Üzerinde Naive Bayes Sınıflandırması Sonucu

Ayrıca Verilerim üzerinden karşılaştırma yapmak amaçlı bazı smile işaretlerini çıkartıp naive bayes classifier uyguladığım zaman **%76** oranında f1-score değerini sağlıyoruz.

```

Clasification:
      precision    recall  f1-score   support

 negative         0.79         0.70         0.74         1500
 positive         0.73         0.82         0.77         1500

 avg / total         0.76         0.76         0.76         3000

Confussion matrix:
[[1043  457]
 [ 272 1228]]

```

Şekil 14 : Bazı Smile İşaretleri olmadan yapılan Naive Bayes Sınıflandırması Sonucu

Şekil-15 üzerinde verilerimiz içerisinde modelimizin etkili bulunduğu 20 adet özelliğimiz gösterilmiştir. Örnek olarak verilerimizde “:)” smile ifadesinin pozitif olarak etiketlenmesi negatif olarak etiketlenmesinden 1005 kat daha fazladır.

```

Most Informative Features
:( = True          negati : positi = 2071.0 : 1.0
:) = True          positi : negati = 1005.4 : 1.0
See = True         positi : negati = 36.3 : 1.0
arrived = True     positi : negati = 32.3 : 1.0
THAT = True        negati : positi = 27.7 : 1.0
miss = True        negati : positi = 26.5 : 1.0
Thank = True       positi : negati = 25.3 : 1.0
x15 = True         negati : positi = 23.7 : 1.0
sad = True         negati : positi = 22.4 : 1.0
Thanks = True      positi : negati = 20.6 : 1.0

```

Şekil 15 : Etkili 10 Özellik

5. SONUÇ

2 Farklı veri seti üzerinden duygu analizi yapmak için naive bayes sınıflandırmasını kullandığımız zaman film yorumları datasetimizin bayes sınıflandırması ile bize sağladığı doğruluk değerimiz **%72** olarak çıkmıştır. Stopwords keimelerini de dahil ettiğimizde bulduğumuz yeni doğruluk değerinin eski doğruluk değerimizden çok farklı olmadığı görülmüştür.

Twitter dataseti üzerinde gerçekleştirmiş olduğumuz naive bayes sınıflandırmasında ise hesapladığımız doğruluk değeri **%99** olarak bulunmuştur. Belirli smile ifadelerini çıkarıp işlem yaptığım zaman doğruluk değerimiz **%76** oranında çıktığı görülmüştür. Buradan twitter verileri üzerinde smile ifadelerinin önemli bir anlamı olduğu görülmüştür.

KAYNAKLAR

Akgül, E. S., Ertano, C., & Diri, B. (2016). Sentiment analysis with Twitter. *Pamukkale University Journal of Engineering Sciences*, 22(2), 106–110.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python. Text* (Vol. 43).

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval. Information Retrieval*.

Parveen, H., & Pandey, S. (2016). Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm, 416–419.

Rana, S., & Singh, A. (2016). Comparative Analysis of Sentiment Orientation Using SVM and Naïve Bayes Techniques. *International Conference on Next Generation Computing Technologies*, (October), 106–111.

Şeker, Ş. E. (2015). *Weka ile Veri*

Madenciliği.

Theodoridis, S., & Koutroumbas, K.
(2009). *Pattern Recognition*.
<https://doi.org/http://www.gbv.de/dms/ilmenau/toc/576990965.PDF>

Türkmenoğlu, C. (2015). Türkçe
Metinlerde Duygu Analizi (Yüksek
Lisans Tezi), 47.

Üran, A. G., Uysal, M., & Doğrusöz, Ö.
(2014). Destek Vektör Makineleri
Parametre Optimizasyonunun Duygu
Analizi Üzerindeki Etkisi. *DEÜ
MÜHENDİSLİK FAKÜLTESİ
MÜHENDİSLİK BİLİMLERİ
DERGİSİ*, 16, 86–93.

Yarrow, K., & Sverdrup-Stueland, I.
(2004). Twitter Predicts Swine flu
outbreak in 2009.
Openaccess.City.Ac.Uk, 47(May),
552–567.