Google n-gram Veri Tabanı ile Cinsiyete Göre Üzüntü ve Mutluluk Analizi

Sadnessand Happiness Analysis by Gender Using Google n-gram Database

İlknur DÖNMEZ
İstanbul Aydin Üniversitesi
ilknurdonmez@aydin.edu.tr

Elena BATTINI SÖNMEZ
İstanbul Bilgi Üniversitesi
elena.battini@bilgi.edu.tr

Öz

"Bilgi çağı" ya da "dijital çağ" olarak adlandırılan 21. yüzyılda hayatımızın her alanında kullandığımız veri, elektronik olarak toplanabilmekte, işlenebilmekte, analiz edilip kullanılabilmektedir. Dijital veriler sosyal ağlardan, kullandığımız araçlardan (Nesnelerin İnterneti), kamera sistemleri ve OCR sitemleri gibi günlük hayatta kullandığımız bilgileri dijital bilgiye çeviren pek çok araç tarafından elde edilebilmektedir. Günümüzde çığ gibi büyüyen büyük verinin analiz edilmesi ve veriyi bilgiye dönüştürecek faydalı kalıpların bulunması önemli bir konudur. Bu çalışmada "Mutluluk" ve "Hüzün" gibi iki temel insan duygusu cinsiyet durumu da dikkate alınarak, Google n-gram derleminden faydalanılarak analiz edilmiştir. Bu derlem, 1500 ve 2008 yılları arasında yayınlanan milyonlarca kitap taranarak elde edilmiştir. İnsanların milyonlarca kitapta kullandığı sözcüklerden oluşan bu derlem, insana özgü özellik ve davranışlar için bir gösterge olarak düşünülebilir. Bu çalışma, insan duygularının, duygularına karşılık gelen sözcüklerin sıklığıyla tahmin edilebileceği hipotezine dayanmaktadır. Makalemizde, gelecek yıllardaki "Mutluluk" ve "Hüzün" duygularının kullanım sıklığını cinsiyet kırılımına göre tahmin etmek için regresyon analiz yöntemleri kullanılmaktadır. Bu çalışma "Google n-gram Veri tabanı ile Üzüntü ve Mutluluk Üzerine Duygu Analizi"çalışmasının cinsiyet kırılımını içeren genişletilmiş halidir.

Gönderme ve kabul tarihi: 16.10.2018-24.11.2018
İ. Dönmez: orcid.org/0000-0002-8110-4084
E. B. Sönmez: orcid.org/0000-0003-0090-984X
Makale türü: Araştırma

***Anahtar Sözcükler**:* Duygu çıkarımı, cinsiyete göre üzüntü ve mutluluk analizi, duygu tahmini, duygu madenciliği, Google n-grams, duygu analizinde cinsiyet kırılımı

Abstract

The currentera has been defined as "Dijital Age" and "Information Age" since it is characterized by an exponential growth of data, generated by bothhuman, i.e. social environments, and machine, i.e. Internet of things. The challenge is to convert "data" into "information", by analyzing the data and discovering patterns hidden inside it. In this paper the two basic human feelings of Happiness and Sadness are extracted from a subset of Google n-grams corpus and analyze daccording to gender. Google n-gramscorpus is generated from millions of scanned boks published between year 1500 and 2008; it can be considered as an indicator forhuman specific feature and behavior. Under the hypothesis thatuser'semotion can be extrapolated by the frequency of the corresponding emotional words, this study applies regression to predict the importance of the Happiness and Sadness emotional states in future years. This study is enhanced version of the "Feeling Analysis for Sadness and Happiness using Google n-gram Database"study which includes gender fraction for each analysis.

***Keywords*:** Feeling extraction, sadness analyses, happiness analyses acording to gender, feeling prediction, feeling mining, Google n-grams, gender fraction for feelings analysis

1. Giriş

Sözcükler önemlidir. İnsanlar her gün iş yerinde, okulda ve günlük hayatlarında iletişim kurmak için sözcükleri kullanırlar. Sözcükler sayesinde insanoğlunun oluşturduğu bilgiler sözlü ya da yazılı olarak nesilden nesle aktarılır. Dilimizdeki "*Söz uçar yazı kalır*" değimi sözcüklerin yazılı haline vurgu yapmaktadır. Dilin (sözcüklerin) yazılı ifadelerinin önemli bir kısmını oluşturan kitaplar; insan bilgisini ve birikimini içerirler. Ayrıca kitaplar, insan davranışlarını, duygularını ve fikirlerini yansıtırlar. Bu çalışmada, üzüntü ve mutluluk üzerine duygu analizi yapmak ve yakın gelecekte bu temel duygulardaki değişimleri tahmin etmek için bu duyguların kitaplardaki ifade ediliş sıklığı kullanılmıştır. Bu amaçla çalışmamızda Google n-gram veri tabanı [1], [2] kullanılmıştır. Google n-gram veri tabanı, 1500 ile 2008 yılları arasında yayınlanan milyonlarca kitabın taranmasından elde edilmiştir. Farklı sözcük ve sözcük gruplarını kitaplarda görülme sıklığına göre grafik olarak gösteren bir arayüze sahiptir [3]. Grafikte "x" ekseni zaman çizgisini ve "y" ekseni aranan sözcük grubunun yüzde cinsinden göreceli sıklığını ifade eder. Bu veri tabanında aranacak sözcük öbeği uzunluğu maksimum beştir.

Şekil-1: Google-n gram arama sonuç arayüzü

Şekil-1'deki Google n-gram arama sayfasında "I feelsad" sözcük grubunun yayın zamanına göre kitaplarda görülme sıklığı grafik olarak gösterilmektedir. Araştırmamızda kitapların toplumu yansıttığı ve kitaplarda kullanılan belirli sözcüklerin sıklığının insan davranış analizi ve tahmini için kullanılabileceği varsayılmıştır. Bu çalışmada "Korku", "Öfke", "İğrenme", "Mutluluk, "Üzüntü" ve "Şaşırma" gibi altı temel duygudan ikisi "Mutluluk" ve "Üzüntü" incelenmiştir.

2000'li yılların başında duygu analizi çalışmaları yapılmaya başlanmıştır ve literatürde bu alanla ilgili pek çok çalıma mevcuttur. Duygu analiz çalışmaları iki grupta incelenebilir. İlk grup, duyargalar (sensör) aracılığıyla duygu durumunu ölçen gerçek zamanlı çalışmalardır. 2004 tarihli bir çalışmada fizyolojik duyargalar ve hareketli metinler kullanılarak internet üzerinden görüşmelerde duygu durumları ölçülmektedir [4]. 2010 yılında yapılan başka bir çalışmada, dinleyicilerin otuz farklı müzik parçasına verdiği duygusal tepkiler (mutluluk ve hüzün) ve bu tepkiler arasındaki benzerlikler ve farklılıklar incelenmiştir [5]. Elektro-ensefalografi (EEG)ya da Beyin Çizgesi Yöntemi ölçümüne dayalı gerçek zamanlı duygu tanıma çalışmasında, duyguların ölçümü EEG sinyalleri aracılığıyla yapılmıştır [6].

İkinci grup metinler üzerinden duyguyu araştıran duygu analiz çalışmalarıdır. 1974 tarihli bir çalışma öznel duyguların sınıflandırılmasında analog ölçekleri kullanmaktadır [7]. 2002 tarihli bir başka makalede, bir internet sohbet ortamındaki giriş metinleri analiz edilerek, metin ile iletilen duygular çıkartılmakta ve ekranda duygu ile uyumlu sembol gösterilmektedir [8]. 2007 yılında yapılan bir başka çalışmada, insanların metin tabanlı iletişim sırasında duygularını nasıl ifade ettikleri ve karşı tarafın duygularını nasıl tespit ettikleri incelemektedir [9]. Bu çalışmaya benzer şekilde, 2011'de, insanların kendilerini ve başkalarını daha iyi anlamalarına yardımcı olmak için dünyadaki duyguları toplamak amaçlı duygusal bir arama motoru oluşturuldu. Bu çalışmada internet sayfaları ve sosyal paylaşım siteleri taranarak "hissediyorum" sözcüğü arandı ve bu sözcüğün kullanıldığı cümleler analiz edildi [10]. 2013 yılında duygu analizi ve görüş madenciliği çalışmasında üzüntü ve mutluluk duygularının analizi yapılmıştır [11].

Google n-gram veri tabanını kullanan farklı çalışmalar vardır. Bu çalışmanın bir tanesi, 2009 yapılmıştır ve terim sayısı ve belge sıklığı arasındaki ilişki ile ilgilidir [12]. 2011 yılındaki başka bir çalışmada bu veri tabanı kullanılarak dil modelleri çıkarılmaktadır [13]. Ayrıca, Google n-gram kullanılarak sözcükler arasındaki ilişki ölçümleri [14], metin benzerliği [15] konuları incelenmiş ve Google n-gram veri tabanı kültürel karmaşıklığı hesaplamak için kullanılmıştır [16]. Google n-gram veri tabanı farklı dil veri kümeleri içerdiğinden, farklı diller için anlamsal benzerliğin bulunmasında kullanılmıştır [17]. Google n-gramların akademik kullanımını kolaylaştırmayı amaçlayan bir çalışma da mevcuttur [18]. Bu çalışmada ise veri kümesini farklı bir mimari ve arayüzle kullanma alternatifleri incelenmiş, Google n-gram veri tabanı ile "Üzüntü ve Mutluluk

Üzerine Duygu Analizi" çalışması **cinsiyet kırılımı bazında** değerlendirilerek tartışılmıştır.

Literatürde duyguların cinsiyet kırılımı üzerine bazı çalışmalar mevcuttur. 2005 yılındaki "Ergenlik dönemindeki bireylerde cinsiyet ve sağlığa göre mutluluk" çalışmasında mutluluk bazı sağlık değişkenlerine bağlanmıştır [19]. 151 kişi ile yapılan bu çalışmada kızlar ve erkekler arasında önemli bir fark bulunmamıştır. 2003 yılındaki, 2199 kişinin katıldığı cinsiyet ve duygular çalışmasında mutluluk ve üzüntünün de bulunduğu sekiz duygu incelenmiştir [20]. İnsanların bu tarz duygulara verdiği tepkiler ölçülmüştür. Bu çalışmada sevinç, korku, hayret, üzüntü, iğrenme, öfke ve beklenti duygularına bayanların verdiği tepki, erkeklerin verdiği tepkilerden fazla olarak ölçülmüştür. Sadece güven duygusunda erkeklerin puanı bayanlardan yüksek bulunmuştur. Bizim çalışmamız Google n gram veri tabanından faydalandığı için diğer çalışmalara göre çok daha büyük bir veri kümesi üzerinde inceleme yapmaktadır. İnsanı ifade eden 100 çeşit anahtar kelime seçilmiştir (teyze, abla, kral, öğretmen vb.). Bu 100 anahtar kelimenin önüne mutlu, aynı 100 anahtar kelimenin önüne üzgün konularak bu ikili sıfat tamlamaları tüm veri tabanında aratılmıştır. Çalışmamız büyük bir veri kümesinde farklı iki duyguyu aynı anahtar sözcüklerle birleştirerek birbiriyle kıyaslanabilir sonuçlar elde etmeyi hedeflemektedir.

2. Veri

Google n-gram veri tabanı Jon Orwant ve Will Brockman tarafından geliştirildi ve Aralık 2010'da yayınlandı. Veri tabanının en yeni sürümü olan 2. sürümü Google tarafından2012 yılında 8.116.746 kitap seçilip taranarak oluşturuldu [2].

Google N-gram veri tabanı, yayınlanan tüm kitapların %6'sına denk olan 8 milyondan fazla kitabın sözcüklerini içerdiğinden, dilbilimsel araştırmalara ve kültürel eğilimlerin analizine olanak sağlar.

Aşağıdaki Çizelge-1, Google n-gram veri tabanının ikinci sürümünde farklı dil grupları için taranan kitap sayısını ve içerilen sözcük sayısını göstermektedir.

Çizelge -1: Google n-gram veri tabanı içeriği

Dil	Kitap Sayısı	Sözcük Sayısı
İngilizce	4.541.627	468.491.999.592
İspanyolca	854.649	83.967.471.303
Fransızca	792.118	102.174.681.393
Almanca	657.991	64.784.628.286
Rusça	591.310	67.137.666.353
İtalyanca	305.763	40.288.810.817
Çince	302.652	26.859.461.025
Musevice	70.636	8.172.543.728

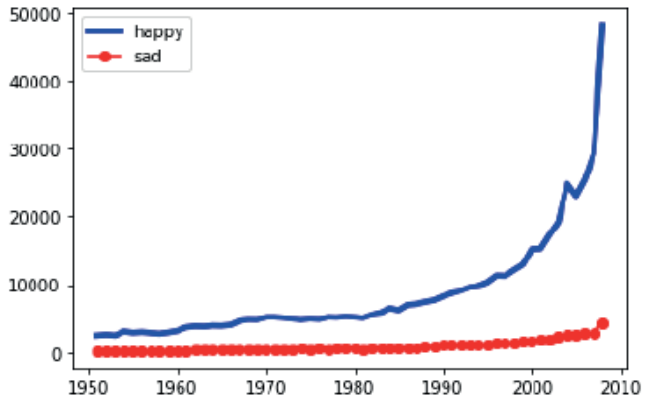
Çalışmamız, Google n gram veri tabanının yalnızca küçük bir alt kümesini kullanmaktadır. Çalışmamızda 2-gramlık İngilizce derlem olan 'googlebooks-eng-all-2gram-20120701-ha' ve 'googlebooks-eng-all-2gram-20120701-sa' veri tabanları kullanılmıştır. Bu veri tabanı ikili sözcük gruplarından oluşmaktadır. Bu veri tabanlarında ilk sözcüğün "happy (mutlu)" ya da "sad (üzgün)" olduğu, ikinci sözcüğün ise insanlar tarafından kullanılabilen bir sözcük olduğu gruplar dikkate alınmıştır. Zaman olarak sadece 1950 yılından sonra yayınlanmış kitaplardaki ifadeler dikkate alınmıştır.

Çizelge-2, Google n-gram veri tabanının çalışmada kullanılan kısmını göstermektedir. Çalışmamızda bu veri tabanının 2- gramlık bölümlerinden 1950 yılından sonraki kayıtlar içinde "happy (mutlu)" ve "sad (üzgün)" sözcüklerinin bir insanı nitelediği sözcük grupları aranarak çıkarılmıştır. Çıkarılan ikili sıfat öbekleri analiz edeceğimiz verileri oluşturmaktadır.

Çizelge-2: Veri tabanının çalışmada kullanılan bölümü

1950'den sonraki yayınlarda "happy (mutlu)" sözcüğünün bir insanı niteler şekilde kullanımı	
Cümle sayısı	501.962
Kitap sayısı	417.234
1950'den sonraki yayınlarda "sad (üzgün)" sözcüğünün bir insanı niteler şekilde kullanımı	
Cümle sayısı	53.234
Kitap sayısı	46.468

Şekil-2 "Sad" ve "Happy" sözcüklerinin insanı niteleyen sözcüklerle birlikte kullanıldığı 2-gram sözcüklerin görülme sıklıklarını grafik olarak göstermektedir. İkili n-gramlar içinde ilk sözcüğün üzgün ve mutlu olduğu sözcükler gruplanmıştır.



Şekil-2: 2-gram içinde "Sad" ve "Happy" sözcüklerinin insanı niteleyen sözcüklerle birlikte kullanım sıklığı

Sonuçlar arasından insan için kullanılan 100 çeşit sözcük rastgele seçilmiştir. Sonuçların anlamlı çıkması için hem mutlu sıfatı hem de üzgün sıfatı aynı 100 çeşit sözcükle birlikte veri tabanı içinde aratılmıştır. Çizelge-3 analiz çalışmamız için kullanılan ifadelerin bazılarını sıklıklarıyla birlikte göstermektedir.

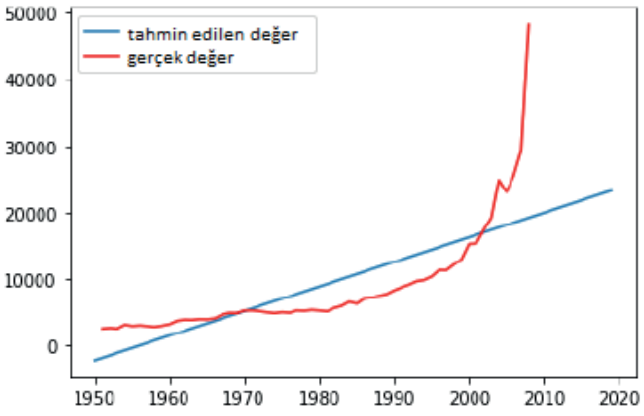
Çizelge -3: Kullanılan veri örnekleri

happy baby	7774	sad child	3069
happy boy	13733	sad boy	2181
happy couple	53013	sad couple	348
happy couples	11922	sad couples	74
happy dad	314	sad dad	398
happy daddy	191	sad daddy	99

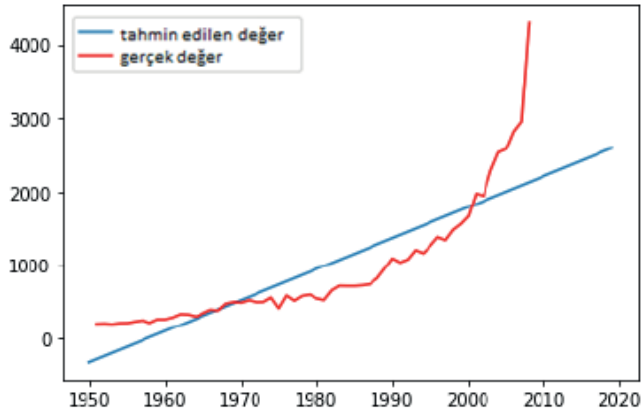
3. Yöntem

Bu çalışmada duygu analizi ve tahmini yapmak için birinci (doğrusal), ikinci ve üçüncü derece regresyon yöntemleri kullanılmaktadır. Tüm kodlama "Python" programlama dili ve "sci-kit" kütüphanesi kullanılarak yazılmıştır. Bu yöntemin amacı veriye en uygun modeli tanımlamak ve bu modeli kullanarak tahmin yapmaktır.

Basit doğrusal regresyon yönteminde, verinin $y = ax + b$ fonksiyonu ile uyumlu olduğu varsayılır. Burada giriş verilerimiz zaman; çıkış verilerimiz ise aranan sözcüklerin görülme sıklığıdır. Öncelikle $y = ax + b$ fonksiyonu için elimizdeki veri kullanılarak toplam hatayı en aza indiren a ve b parametreleri bulunmaktır. Teknik olarak, a parametresi, çizginin eğimi ve b, kesişme olarak adlandırılır, çünkü b fonksiyon eğrisinin y eksenini kestiği noktadır. Makine öğrenimi terminolojisinde, bu her tahmin için eklenecek sapmayı (bias) ifade eder. Şekil 3 ve 4'te, a ve b değeri bulunarak elde edilen modeller orijinal verilerle birlikte gösterilmektedir. X ekseni 1950 ile 2020 arasındaki tarihi, y ekseni göreceli frekansı göstermektedir.



Şekil-3: "Happy ..." sözcük grubu için birinci derece regresyon eğrisi

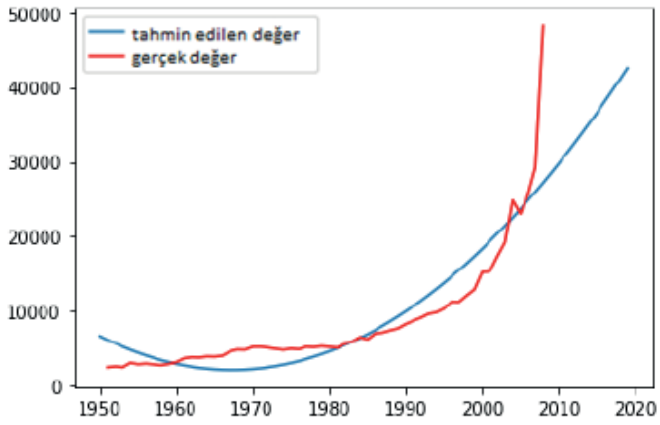


Şekil-4: "Sad ..." sözcük grubu için birinci derece regresyon eğrisi

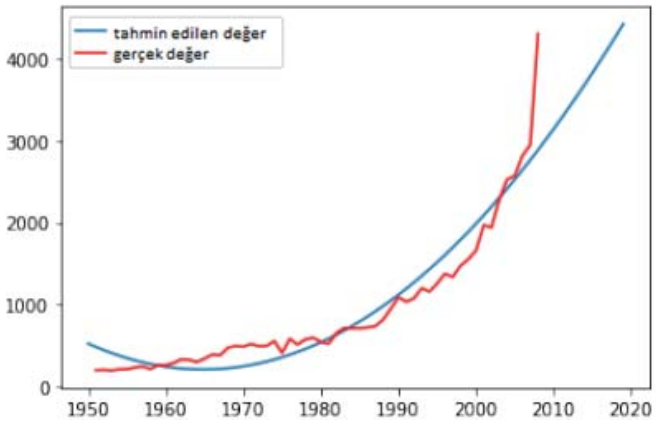
İkinci mertebeden regresyon için, elimizdeki verilerin ikinci mertebeden bir polinom $y = ax^2 + bx + c$ modeline uygun olduğu varsayılır ve bu modeldeki a, b ve c katsayıları elimizdeki (x,y) örneklerini kullanarak hatayı en az yapacak şekilde hesaplanır.

Benzer şekilde üçüncü mertebeden regresyon için modelimizin $y = ax^3 + bx^2 + cx + d$ formunda olduğu var sayılır ve bu modeldeki a, b, c ve d katsayıları elimizdeki (x,y) örneklerini kullanarak hatayı en az yapacak şekilde hesaplanır.

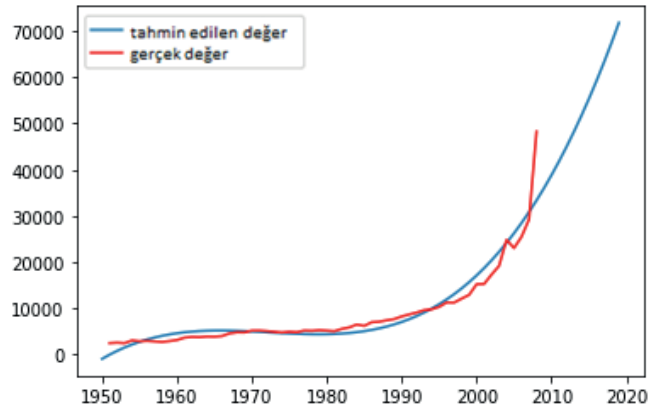
Şekil 5, 6, 7 ve 8, normalleştirilmiş verileri önerdiğimiz modellerle birlikte göstermektedir. Tüm grafiklerde "mutlu" ve "üzgün" veri kümeleri ayrı ayrı incelenmiştir.



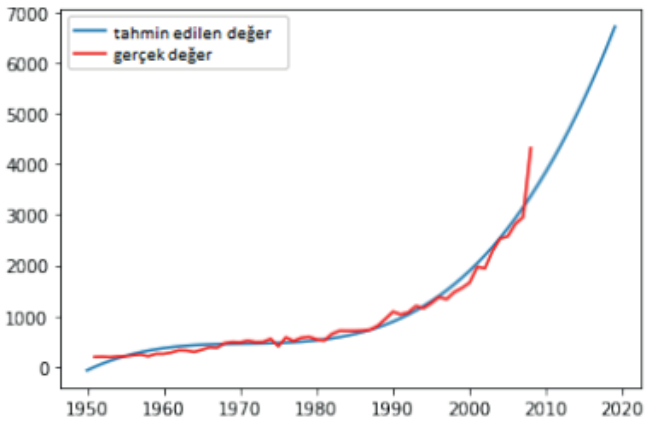
Şekil-5: "Happy ..." sözcük grubu için ikinci derece regresyon eğrisi



Şekil-6:"Sad ..." sözcük grubu için ikinci derece regresyon eğrisi



Şekil-7: "Happy ..." sözcük grubu için üçüncü derece regresyon eğrisi



Şekil-8: "Sad ..." sözcük grubu için üçüncü derece regresyon eğrisi

Çalışmamızda ikili sözcüklerden oluşan veriler yayınlandıkları yıla göre gruplanıp toplam frekansları hesaplandı. Tüm deneylerde zaman ve frekans olarak iki boyut kullanıldı. Zaman boyutu, yayın tarihine karşılık gelirken; frekans boyutu, aranan sözcük grubunun görülme sıklığı ile ilgilidir. Çalışmamızda iki boyut içinde veriler normalleştirildi.

Hatayı ölçmek için istatistiksel modeli değerlendiren "python" programındaki hazır fonksiyonlar kullanıldı. Elde ettiğimiz üç model için hatayı ölçtüğümüzde, Çizelge-4'de görüldüğü üzere ikinci dereceden regresyon tekniğinin en başarılı olduğu görülmüştür.

Çizelge-4: Modellerin ortalama standart hata değerleri

	Ort. Standart Hata
Üzgün 1. derece	0,148
Üzgün 2. derece	0,108
Üzgün 3. derece	0,269
Mutlu 1. derece	0,153
Mutlu 2. derece	0,105
Mutlu 3. derece	0,364

Çizelge-5 Üzgün ve Mutlu cümlelerinin elde edilen üç modele göre 2010 ve 2020 yıllarındaki frekans tahminlerini içermektedir.

Çizelge-5: Modellere göre tahmin değerleri

Mutlu sayısı	2010	2020
1.derece	19931,76	23629,21
2.derece	29746,46	44164,58
3.derece	29893,61	44615,23
Üzgün sayısı	2010	2020
1.derece	2208,84	2632,13
2.derece	3145,35	4591,59
3.derece	3158,49	4632,66

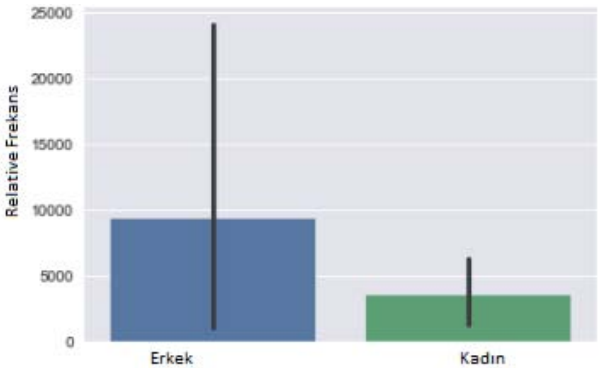
Çizge-5'in sonuçları özetlendiğinde; "Mutlu" içeren sözcük grupları için birinci, ikinci ve üçüncü dereceden modellerin (2010, 2020) tarihleri için öngördüğü görülme sıklığı sırasıyla (19931, 23629), (29746, 44164) ve (29894, 44615)'dir.

Benzer şekilde "Üzgün" içeren sözcük grupları için birinci, ikinci ve üçüncü dereceden modellerin (2010, 2020) tarihleri için öngördüğü görülme sıklığı sırasıyla (2209, 2632), (3145, 4592) ve (3158, 4633)'dür. 2. mertebeden regresyon modelimiz en az hataya sahip olduğundan üretilen tahmin değerlerinden 2. modelin önerdiği tahmin (3145, 4592) diğerlerinden daha doğrudur.

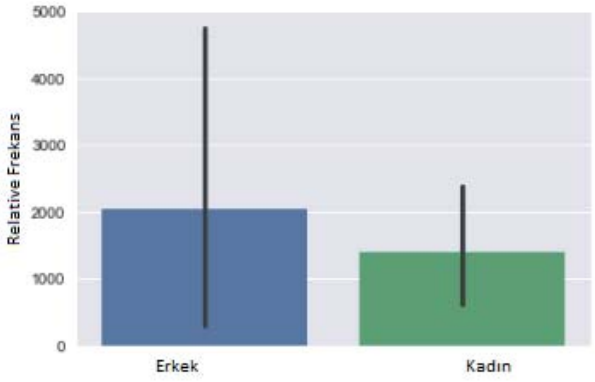
4. Cinsiyete göre duygu analizi

Bu çözümlemenin anlamlı sonuç vermesi için insanı ifade eden100 anahtar kelime seçildi (Kral, anne, kraliçe, insan gibi). Bu anahtar sözcükler "bayan", "erkek" ve "cinsiyetsiz" olarak elle işaretlendi. Zaten bu gruplar yaklaşık eşit sayıda seçilmişti (35-35-30). Bayan ve erkeği ifade eden anahtar sözcükler hem önüne "Happy (mutlu)"; hem de önüne "Sad (üzgün)" getirilerek tüm veri tabanında aratıldı. Örnek olarak mutlu insan, mutlu anne, mutlu baba, üzgün kral, üzgün çift ve üzgün teyze gibi anahtar sözcükler bayan için söylenmişse "bayan"; erkek için söylenmişse "erkek" olarak gösterildi.

İçinde mutlu olan sözcük grupları erkek ve bayanın mutluluğundan bahsetmesine göre ayrıştırıldı. Aynı şekilde içinde üzgün olan sözcük grupları erkek ve bayanın üzgünlüğünden bahsetmesine göre ayrıştırıldı. Sonuçlar kıyaslandı.
Şekil-9 ve Şekil-10 ayrı ayrı 1950-2020 yayınları arasında mutlu ve üzgün sözcüklerini içeren ve insanı niteleyen sözcük gruplarının cinsiyete göre kırılımlarını içermektedir. Çizimler ortalama değerleri göstermektedir. Erkek ve kadınlar için çubuk gösterim üzerindeki ince dikey çizgiler %95 güven aralığına göre ortalama değerden sapmayı göstermektedir.



Şekil-9: "Mutlu" sözcüğünün insanı nitelediği sözcük grupları arasında cinsiyet kırılımı



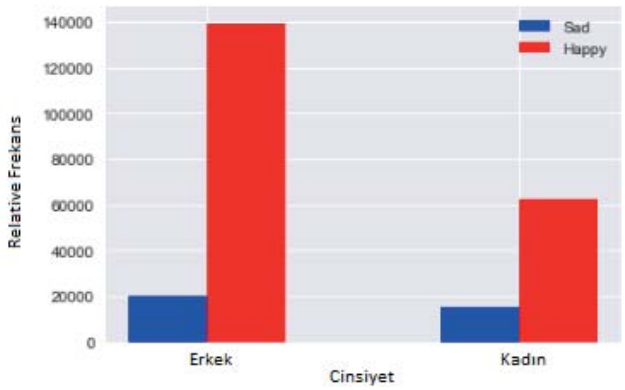
Şekil-10: "Üzgün" sözcüğünün insanı nitelediği sözcük grupları arasında cinsiyet kırılımı

Çizelge-6: İkili sözcüklerin analizi

	Maksimum	Ortalama	Toplam
Mutlu Erkek	108.424	9.298,13	139.472
Mutlu Kadın	19.572	3.463,22	62.338
Üzgün Erkek	13.760	2.032,20	20.322
Üzgün Kadın	5.356	1.395,54	15.351

Çizelge-6'da erkek için kullanılan ikili sözcüklerin en yüksek değerinin ne olduğu (Maksimum), ikili sözcük frekanslarının ortalama değeri ve erkeği ifade eden tüm sözcüklerin toplam karşılaşılma sıklığı mutlu ve üzgün için ayrı ayrı gösterilmiştir. Maksimum değerine bir örnek vermek gerekirse; "Happy man" ikili sözcüğü 1950-2020 yayınları arasında en çok rastlanan ikili sözcüktür ve karşılaşılma frekansı 108.424'dir.

Şekil-11'de mutlu ve üzgün sözcüklerini içeren ve insanı niteleyen sözcük gruplarının cinsiyete göre kırılım değerleri aynı ölçekte gösterilmektedir.Bu grafikte ortalama değerler değil toplam görülme sıklıkları ele alınmıştır. Şekilden görüldüğü üzere araştırmamızda mutlu olarak ifade edilen erkek sayısı, üzgün olarak ifade edilen erkek sayısının neredeyse 7 katıdır. Kadınlar için ise mutlu olarak ifade edilen kadın sayısı, üzgün olarak ifade edilen kadın sayısının yaklaşık 3 katıdır. Kadın ve erkeklerin yaklaşık eşit sayıda mutsuzluklarından bahsedilirken, mutlu olarak bahsedilen erkek sayısı kadın sayısının 2 katından fazladır.



Şekil-11: 1950-2020 yayınları arasında "Mutlu" ve "Üzgün" sözcük gruplarının cinsiyet kırılımı

5. Sonuç

Çalışmamız, kitapların insan duygularını ifade ettiği hipotezine dayanmaktadır. Bu hipoteze dayanarak kitaplar içinde mutlu ve üzgün sözcüklerinin bir insanı niteler biçimde kullanıldığı sözcük grupları giriş verisi olarak kullanılmaktadır. Çalışmamızda bu sözcüklerin zaman içindeki frekansları incelenmiştir. Elimizdeki verilere en uygun modeller önerilmiş ve bu modeller ile yakın gelecekte (2020) bu duyguların ifade edilme frekansı tahmin edilmiştir. Deneyler sonucunda ikinci dereceden regresyonun en az hatayı verdiği görülmüştür. Sonuçta mutlu sözcüğünün 2020 yılında görülme sıklığı 44164; üzgün sözcüğünün 2020 yılında görülme sıklığı 4592 olarak tahmin edilmiştir. Son olarak mutluluk ve hüznü ifade eden sözcük gruplarının hangi cinsiyetle ne kadar kullanıldığı incelenmiştir. Bu inceleme sonucunda kitaplarda mutlu sözcüğünün daha çok erkekleri niteler şekilde kullanıldığı, mutlu sözcüğünün erkekleri niteler şekilde kullanımının bayanları niteler şekilde kullanımında neredeyse 7 kat fazla olduğu ve üzgün sözcüğünün kadın ve erkekler için yaklaşık olarak eşit kullanıldığı görülmüştür.

6. Kaynakça

[1] Michel, J. B et all, (2011). The Google Books Team, 176-182.

[2] Michel, J. B. et all, (2011). Quantitative analysis of culture using millions of digitized books. science, 331(6014), 176-182.

[3] Smallwood, C.,(2015). The complete guide to using Google in libraries: instruction, administration, and staff productivity (Vol. 1). Rowman & Littlefield.

[4] Wang, H., Prendinger, H., & Igarashi, T. (2004, April). Communicating emotions in online chat using physiological sensors and animated text. In CHI'04 extended abstracts on Human factors in computing systems (pp. 1171-1174). ACM.

[5]Hunter, P. G., Schellenberg, E. G., & Schimmack, U. (2010). Feelings and perceptions of happiness and sadness induced by music: Similarities, differences, and mixed emotions. Psychology of Aesthetics, Creativity, and the Arts, 4(1), 47.

[6] Liu, Y., Sourina, O., & Nguyen, M. K. (2011). Real-time EEG-based emotion recognition and its applications. In Transactions on computational science XII (pp. 256-277). Springer, Berlin, Heidelberg.

[7] Bond, A., & Lader, M. (1974). The use of analogue scales in rating subjective feelings. British Journal of Medical Psychology, 47(3), 211-218.

[8] Zhe, X., & Boucouvalas, A. C. (2002, July). Text-to-emotion engine for real time internet communication. In Proceedings of International Symposium on Communication Systems, Networks and DSPs (pp. 164-168).

[9] Hancock, J. T., Landrigan, C., & Silver, C. (2007, April). Expressing emotion in text-based communication. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 929-932). ACM.

[10] Kamvar, S. D., & Harris, J. (2011, February). We feel fine and searching the emotional web. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 117-126). ACM.

[11] Kaur, A., & Gupta, V. (2013). A survey on sentiment analysis and opinion mining

techniques. Journal of Emerging Technologies in Web Intelligence, 5(4), 367-371.
[12] Klein, M., & Nelson, M. L. (2009, April). Correlation of term count and document frequency for Google n-grams. In European Conference on Information Retrieval (pp. 620-627). Springer, Berlin, Heidelberg.
[13] Pauls, A., & Klein, D. (2011, June). Faster and smaller n-gram language models. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 pp. 258-267
[14] Islam, A., Milios, E., & Keselj, V. (2012). Comparing Word Relatedness Measures Based on Google n grams. Proceedings of COLING 2012: Posters, 495-506.
[15] Islam, A., Milios, E., & Kešelj, V. (2012, May). Text similarity using google tri-grams. In Canadian Conference on Artificial Intelligence (pp. 312-317). Springer, Berlin, Heidelberg.
[16] Juola, P. (2013). Using the Google N-Gram corpus to measure cultural complexity. Literary and linguistic computing, 28(4), 668-675.
[17] Joubarne, C., & Inkpen, D. (2011, May). Comparison of semantic similarity for different languages using the Google N-gram corpus and second-order co-occurrence measures. In Canadian Conference on Artificial Intelligence (pp. 216-221). Springer, Berlin, Heidelberg.
[18]Davies, M. (2014). Making Google Books n-grams useful for a wide range of research on language change. International Journal of Corpus Linguistics, 19(3), 401-416.
[19]Mahon, N. E., Yarcheski, A., & Yarcheski, T. J. (2005). Happiness as related to gender and health in early adolescents. Clinical nursing research, 14(2), 175-190.
[20]Brebner, J. (2003). Gender and emotions. Personality and individual differences, 34(3), 387-394.