**T.C.**

**HARRAN ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ**

# YÜKSEK LİSANS TEZİ

**VERİ MADENCİLİĞİ TEKNİKLERİ KULLANILARAK SOSYAL MEDYA VERİLERİ İLE DUYGU ANALİZİ**

**Ayhan AKKAYA**

**BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

**ŞANLIURFA 2022**

**İÇİNDEKİLER**

**Sayfa No**

[ÖZET ii](#_bookmark0)

[ABSTRACT iii](#_bookmark1)

[TEŞEKKÜR iv](#_bookmark2)

[ŞEKİLLER DİZİNİ v](#_bookmark3)

[ÇİZELGELER DİZİNİ vi](#_bookmark4)

[SİMGELER VE KISALTMALAR DİZİNİ vii](#_bookmark5)

1. [GİRİŞ 1](#_bookmark6)
   1. [Çalışmanın İçeriği 7](#_bookmark18)
2. [ÖNCEKİ ÇALIŞMALAR 9](#_bookmark19)
3. [MATERYAL ve YÖNTEM 13](#_bookmark20)
   1. [Materyal 13](#_bookmark21)
      1. [Veri Toplama 13](#_bookmark22)
         1. Reddit 14
      2. [Veri Temizleme 13](#_bookmark22)
      3. [Veri Depolama 13](#_bookmark22)
   2. Yöntem 14
      1. Verilerin Etiketlenmesi 15
         1. 3 Sınıflı Duygu Etiketleme 14
         2. 4 Sınıflı Duygu Etiketleme 14
      2. Verilerin Eğitilmesi 15
         1. Kelime Gömme 16
         2. Bert 16
4. ARAŞTIRMA BULGULARI VE TARTIŞMA 17
   1. Değerlendirme Metrikleri 18
   2. 3 Sınıflı Duygu Eğitim Sonuçları 19
   3. 4 Sınıflı Duygu Eğitim Sonuçları 19
   4. Aşı Analizi 20
5. [SONUÇLAR ve ÖNERİLER 44](#_bookmark74)

[KAYNAKLAR 45](#_bookmark75)

ÖZGEÇMİŞ 49

**Yüksek Lisans Tezi**

**VERİ MADENCİLİĞİ TEKNİKLERİ KULLANILARAK SOSYAL MEDYA VERİLERİ İLE DUYGU ANALİZİ**

**Ayhan AKKAYA Harran Üniversitesi**

**Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı**

**Danışman: Dr. Öğr. Üyesi Nagehan İLHAN**

**YIL: 2022, Sayfa: ---**

Sosyal medyanın gittikçe daha da önem kazandığı bu zamanda birçok konuda adeta karar organı haline geldiği görülmektedir. Özellikle 2019 yılında hayatımızı etkisi altına alan koronavirüs salgını insanları toplumsal anlamda birbirinden daha da uzaklaştırmıştır. Salgının seyri ile ilgili oluşan her gelişme insanların en önemli merak konusu olmuştur. Bu süreçte salgına son vereceği düşünülen aşılar ile ilgili insanların neler düşündüğü, nasıl bir duygu halinde oldukları kesinlikle bilinmesi gereken bir konu haline gelmişti. Bunun belki de en büyük ispatı salgının başında bu konuda açılan konu başlıklarının halen güncelliğini korumasıdır. Duyguların elde edilmesi kadar bu duyguların doğru anlamlandırılması da önemlidir. Bizde çalışmamızda bu düşünceden yola çıkarak öncelikle insanların en çok yoğunlaştığı konu başlıklarından 2020 ve 2021 yılındaki verileri çektik. Verilerin doğru duygu etiketlerine sahip olması için veriler üzerinde birçok veri ön işleme işlemleri gerçekleştirdik. Çok sınıflı duygu etiketlemesi yapmak için birden fazla modeli üst üste kullanarak hibrit bir model oluşturmaya çalıştık. Duyguları hem pozitif, negatif, nötr ve aynı zamanda korku, mutsuzluk, eğlenceli, mutlu şeklinde farklı sınıflandırıcıları kullanarak etiketledik. %93 gibi bir doğruluk oranı elde ettik. Bu yüksek doğruluk oranına inanarak insanların aşı ve aşı markaları ile ilgili düşüncelerine odaklandık. Çalışmalarımızın doğruluğunu arttırmak için derin öğrenme tabanlı sınıflandırıcıları kullanmaya çalıştık. Bu çalışmalarımızın duygu analizi ile ilgili farklı alanlarda da başarılı olacağını düşünmekteyiz. Yorumların artması büyük verinin ortaya çıkmasına neden olmaktadır. Bizde sonraki çalışmalarımızda büyük veri teknolojisini kullanarak çalışmamızın kapsamını geliştirmek istiyoruz.

**ANAHTAR KELİMELER:** Duygu Analizi, Sosyal Medya, Aşı, Koronavirüs, Doğal Dil İşleme

.

**MSc Thesis**

**EMOTION ANALYSIS WITH SOCIAL MEDIA DATA USING DATA MINING TECHNIQUES**

**Ayhan AKKAYA**

**Harran University**

**Graduate School of Natural and Applied Sciences**

**Department of Computer Engineering**

**Supervisor: Assist. Prof. Dr. Nagehan İLHAN**

**Year: 2022, Page No: ---**

At this time when social media is gaining more and more importance, it is seen that it has almost become a decision-making body on many issues. Especially in 2019, the coronavirus epidemic, which affected our lives, has distanced people from each other in a social sense. Every development related to the course of the epidemic has been the most important curiosity of people. In this process, what people thought and how they felt about the vaccines, which were thought to end the epidemic, became a subject that should definitely be known. Perhaps the biggest proof of this is that the topic titles opened at the beginning of the epidemic are still up-to-date. It is important to understand these emotions correctly as well as to obtain them. In our study, based on this idea, we first drew the data in 2020 and 2021 from the topics that people concentrated on the most. We performed a lot of data preprocessing on the data so that the data has the correct sentiment labels. We tried to create a hybrid model by using multiple models on top of each other to do multi-class emotion labeling. We labeled emotions using different classifiers as positive, negative, neutral as well as fear, unhappiness, fun, happy. We achieved an accuracy rate of 93%. Believing this high accuracy rate, we focused on what people think about vaccines and vaccine brands. We tried to use deep learning-based classifiers to increase the accuracy of our work. We think that these studies will be successful in different areas related to sentiment analysis. The increase in comments leads to the emergence of big data. We want to improve the scope of our work by using big data technology in our next work.

**KEY WORDS:** Sentiment Analysis, Social Media, Vaccine, Coronavirus, Natural Language Processing

# TEŞEKKÜR

Lisansüstü eğitimim boyunca bilgi ve birikimlerini benimle paylaşıp birçok konuda bana destek olan değerli hocam Dr. Öğr. Üyesi Nagehan İLHAN’a, sadece akademik alanda değil hayatın her kademesinde bana sürekli destek olan, zor zamanlarımda sabır gösterip her daim inançları ile güç veren aileme teşekkür ederim.

**Sayfa No**

[Şekil 1.1. 1](#_bookmark8)

[Şekil 1.2. 4](#_bookmark12)

[Şekil 1.3. 4](#_bookmark13)

[Şekil 1.4. 6](#_bookmark14)

[Şekil 3.2. 14](#_bookmark24)

[Şekil 3.3. 15](#_bookmark27)

[Şekil 3.4. 16](#_bookmark28)

[Şekil 3.5. 16](#_bookmark29)

[Şekil 3.6. 17](#_bookmark30)

[Şekil 3.7. 17](#_bookmark31)

[Şekil 3.8. 18](#_bookmark32)

[Şekil 3.9. 20](#_bookmark37)

[Şekil 3.10. 21](#_bookmark39)

[Şekil 3.11. 21](#_bookmark40)

[Şekil 3.12. 22](#_bookmark42)

[Şekil 3.14. 26](#_bookmark47)

[Şekil 3.15. 27](#_bookmark49)

[Şekil 3.16. 28](#_bookmark51)

[Şekil 4.1. 31](#_bookmark56)

[Şekil 4.2. 32](#_bookmark57)

[Şekil 4.3. 32](#_bookmark58)

[Şekil 4.4. 34](#_bookmark60)

[Şekil 4.5. 34](#_bookmark61)

[Şekil 4.6. 35](#_bookmark62)

[Şekil 4.7. 38](#_bookmark66)

[Şekil 4.8. 38](#_bookmark67)

[Şekil 4.9. 39](#_bookmark68)

[Şekil 4.10. 40](#_bookmark70)

[Şekil 4.11. 40](#_bookmark71)

[Şekil 4.12. 41](#_bookmark72)

**Sayfa No**

[Çizelge 1.1. 7](#_bookmark16)

[Çizelge 3.1. 19](#_bookmark35)

[Çizelge 3.2. 20](#_bookmark36)

[Çizelge 3.3. 25](#_bookmark45)

[Çizelge 4.1. 30](#_bookmark54)

[Çizelge 4.2. 31](#_bookmark55)

[Çizelge 4.3. 33](#_bookmark59)

[Çizelge 4.4. 36](#_bookmark63)

[Çizelge 4.5. 37](#_bookmark65)

[Çizelge 4.6. 39](#_bookmark69)





|  |  |
| --- | --- |
| VADER | Valence Aware Dictionary for Sentiment Reasoning |
| GLOVE | Global Vectors for Word Representation |
| NLP | Natural Language Processing |
| BERT | Bidirectional Encoder Representations from Transformers |
| POST | Parts Of Speech Tagging |
|  |  |
|  |  |
|  |  |
|  |  |

# GİRİŞ

İnternetin ve teknolojinin gelişmesiyle birlikte insanlar arasında etkileşim daha da artmıştır. Artan etkileşim beraberinde çok büyük verilerin ortaya çıkmasına neden olmuştur. Veriler çok büyük ve düzensiz olduğu için anlamlı bir hale getirilmesi ihtiyacı doğmuştur. Bu anlamda veriler üzerinde işlemler yapılabilmesi için veri madenciliği teknikleri kullanılmaya başlanmıştır.

Veri madenciliği, makine öğrenimi, istatistik ve veritabanı sistemlerinin kesişme noktasında yöntemleri içeren büyük veri kümelerindeki kalıpları keşfetme sürecidir(Hand, 2007). Veri madenciliği teknikleri, sağlık, ekonomi, spor, siyaset vb. farklı alanlarda kullanılmaktadır. Özellikle sağlık alanında kullanılan veri madenciliği teknikleri çok büyük veriler üzerinde anlamlı hale getirip makine öğrenimi, derin öğrenme tekniklerini kullanarak hastalık takibi ve tespiti, semptomlar, bulgular hakkında öğretici bilgiler ortaya koymaktadır.



Sosyal medyada veri madenciliği teknikleri gün geçtikçe daha çok kullanılır hale gelmiştir. Sosyal medyada insanların konuşmaları, yorumları, resimleri vb. etkileşimleri çok büyük miktarda verilerin ortaya çıkmasına neden olmaktadır. Ayrıca sosyal medya özellikle günümüzde toplumu ilgilendiren birçok olayın gündeme gelmesine ve bazılarının da çözülmesine olanak sağlayan güçlü bir platform olmuştur.

Reddit, bir Amerikan sosyal haber toplama, web içeriği derecelendirme ve tartışma sitesidir. Kayıtlı üyeler, siteye bağlantılar, metin gönderileri ve resimler gibi içerikleri gönderir ve bunlar daha sonra diğer üyeler tarafından yukarı veya aşağı oylanır(Blackburn, 2020). Reddit, her bir konu ile ilgili olan subreddit’ler sayesinde insanların belirli forumlarda bir araya gelmelerine olanak sağlamaktadır. Ayrıca reddit, bu gruplardaki verilerin farklı alanlarda kullanılabilmesi için verilerin çekilebilmesine olanak sağlamıştır. Hatta bununla ilgili uygulamalar ve yazılım dillerine ait kütüphaneler bile yazılmıştır.

Koronavirüs hastalığı (COVID-19), yeni keşfedilen bir koronavirüsün neden olduğu bulaşıcı bir hastalıktır(SINGHAL, 2020). 2019 yılında Çin’in Wuhan şehrinde bir balık pazarında ortaya çıktığı düşünülen bu hastalık başlangıçta o bölgede etkili olsa da ilerleyen süreçte tüm dünyayı etkisi altına alan bir pandemi haline gelmiştir. Artan vaka sayıları ve ölüm oranlarındaki artış ülkelerin çok ciddi kısıtlamalara gitmelerine neden olmuştur. Sokağa çıkma yasağı, okulların kapatılması, alışveriş merkezleri, lokantalar, kafeler ve farklı iş alanlarının belirli şartlar altında çalışma koşulları sosyal hayatı adeta durma noktasına getirmiştir. Tüm bu gelişmelerin yaşanması sosyal medya platformunu daha da etkili hale getirmiştir. Sosyal medya artık sohbet ortamı olmaktan ziyade birçok alanda karar organı haline gelmiştir.

Reddit sosyal medya platformunda her alanla ilgili konular konuşulmaktadır. Bu konular ilgi alanlarına göre farklı subredditler olarak ayrılmıştır. Covid-19 pandemisi sosyal medyanın en çok konuştuğu konulardan biri olmuştur. r/Coronavirus, r/COVID-19 gibi subredditler virüsün etkili olduğu ilk anlardan itibaren etkili bir şekilde kullanılmıştır.

İnsanların virüs ile mücadelede en büyük beklentileri şüphesiz ki aşının bulunması olmuştur. Aşı ile ilgili farklı ülkelere ait çalışmalar faz 3, faz 2 ve faz 1 olarak farklı seviyelerde halen devam etmektedir. Aşının virüsten korumada ne kadar etkili olduğu, yan etkilerinin bulunup bulunmadığı gibi sorular özellikle sosyal medyada sürekli önemli bir gündem maddesi olmuştur. Bu anlamda insanların aşı ile ilgili duygu ve düşünceleri önem kazanmıştır. Farklı sosyal medya platformlarında konuşulan bu konu insanların bu konuya ne kadar ilgi gösterdiğini gözler önüne sermektedir. İnsanlar aşının virüsten koruması kadar farklı amaçlara da hizmet edip etmediğini merak etmeye başlamışlardır. Bu anlamda en etkili alanlardan biri olan sosyal medya platformunda aşı ile ilgili ne kadar çok haber, yorum vs. varsa katılmaya çalışmaktadırlar.

Duygu ölçümlerinin öğrenmeye nasıl içgörü sağlayabileceğini daha iyi anlamak için, bu tür ölçümlerin duygu teorisiyle nasıl ilişkili olduğunu düşünmek önemlidir (Weidman, 2016). Başka bir çalışmada ise, araştırmacıların öfke ve korku gibi duyguları ölçtüğünde, bazen öfke ve korkunun her ikisinin de olumsuz olduğunu düşünerek ölçümlerini değerlik (olumludan olumsuza) boyutuna yeniden kodladıklarına dair gözlemler yapılmıştır (Calvo, 2010).

Duyguları sınıflandırmak için farklı yaklaşımlar kullanılmaktadır. Sözlük tabanlı yaklaşım, sınıflandırma sürecini belirli anlamsal yönelime doğru yönlendirebilecek sözcükleri veya tümcecikleri çıkarmaya odaklanmaktadır. Her gösterge sözcüğünün, duygu sözcük sözlüğünün bazı araçları kullanılarak çıkarılan kendi anlamsal değeri vardır. Genel incelemenin polaritesi, kelimelerinin anlamsal değerlerinin ortalama bir toplamı olarak hesaplanır. Makine öğrenimi, önceki örneklerden öğrenmek ve modeli yeni örneklere uygulamak ve sonuçları izlemek için bilgisayar programlarını eğitmekle ilgilenen bilimdir. Makine öğrenimi farklı modlarda eğitilebilir ve uygulanabilir. İlk öğrenme modu, etiketli vakaların bilgisayar modeline veya öğrenciye üzerlerindeki öğrenme noktasını ayarlaması için sunulduğu ve ardından modelin görünmeyen veriler üzerinde genelleme yapma yeteneğini test etmek için görünmeyen noktaların sunulduğu denetimli öğrenmedir. Denetimsiz öğrenme, öğrenen modelinin, öğrenci değerlendirmesini zorlaştıran etiketler olmadan benzer veri noktalarını birlikte gruplandırmaya çalıştığı bir başka öğrenme modudur. Üçüncü öğrenme modu olan pekiştirmeli öğrenmede öğrenme kavramı oldukça farklıdır. Eğitim verilerinin yerini, kendi programlanmış kurallarını izleyerek ortamıyla etkileşime giren aracı kavramı alır. Etmen, bilgi tabanını önceki deneyimlerinden gelen geri bildirimlerle güncelleyerek deneme yanılma sürecinde öğrenir (Deng, 2014).

# Çalışmanın İçeriği

Biz bu çalışmamızda reddit sosyal medya platformunu kullanarak 2020 yılı başlarından itibaren sosyal medyanın aşı ve aşı markaları hakkındaki verileri sosyal

medya platformu olan redditten çektik. Verileri aylık olarak çekmek için redditten veri çekmeye olanak sağlayan python dilinde yazılmış psaw api kullandık.Reddit platformunda çekilen yorumlardan İngilizce yorumlar üzerinde duyguları etiketlemeye çalıştık. 3 sınıflı (positive-negative-neutral) duygu etiketleme işlemi için Valence Aware Dictionary for Sentiment Reasoning (VADER) sınıflandırıcıyı kullandık. 4 Sınıflı (fear-sadness-joy-surprise) duygu etiketleme işlemi için sparkNLP sınıflandırıcısını kullandık. Daha sonra verilerin doğruluğunu arttırmak için hem 3 sınıflı hem de 4 sınıflı veriler üzerinde kelime temsillerini ifade eden kelime gömme yapısı olan Global Vectors for Word Representation (Glove) yapısını kullandık. Glove ile eğitim işlemleri sonucunda 3 sınıflı veriler için %75 ve 4 sınıflı veriler için %70 gibi bir doğruluk oranını elde ettik. Daha yüksek bir doğruluk oranına ulaşmak için Glove ile eğitilen veriler üzerinde son zamanlarda özellikle Doğal Dil İşleme (NLP) alanında kullanılan Derin Öğrenme tabanlı bir kelime gömme yapısı olan Bidirectional Encoder Representations from Transformers (BERT) ile eğitim işlemi gerçekleştirdik. Bert ile eğitim işlemleri sonucunda 3 sınıflı veriler için %83 ve 4 sınıflı veriler için % 76 gibi bir başarı oranı yakaladık. Ham olarak çekilen veriler üzerinde yapılan bir dizi işlem sonucunda verilerimiz daha anlamlı bir hale gelmiş oldu. Çalışmada odaklandığımız asıl amacımız farklı markalara ve aşıya ait insanların duygularını analiz edip insanlara somut bir çalışma sunmaktı.

# ÖNCEKİ ÇALIŞMALAR

Duygu analizi ile ilgili farklı alanlarda yapılmış birçok çalışma bulunmaktadır.

NLP, bilgisayarların insan dillerinde yazılmış ifadeleri veya kelimeleri anlamasını sağlamaya adanmış bir Yapay Zeka ve Dilbilim alanıdır (Khurana, 2017). NLP'nin araştırılan ve bu tezle ilgili görevlerinden bazıları Adlandırılmış Varlık Tanıma, Konuşma Parçaları Etiketleme, Parçalama, Stopwords kaldırma, Lemmatization vb. metindeki öğeler özel adlarla ilgilidir. Parts Of Speech Tagging (POST) benzer şekilde bir cümleyi girdi olarak alır ve içinde bulunan her kelime için konuşmanın bir bölümünü belirler. NLP'deki bir diğer önemli görev ise Parçalama'dır. Parçalama, cümleleri kelimelere ve her bir kelimeye kendi etiketi veya İsim Cümlesi ve Fiil Cümlesi (NP, VP) gibi sözdizimsel bağıntılı anahtar kelimelerle bölümlere ayırmaya izin verir. Stopwords kaldırma, NLP'de sağlanan anahtar kelimelerin belgeleri işlemeden önce kaldırılması gereken başka bir görevdir. Ancak belgeler veya ifadeler, model tarafından tespit edilmesi gereken tek bir kelime için farklı formlar kullanabilir. Bu, Lemmatization görevi kullanılarak yapılabilir. Lemmatizasyon genellikle, normalde yalnızca çekim sonlarını kaldırmayı ve lemma olarak bilinen bir kelimenin temel veya sözlük biçimini döndürmeyi amaçlayan bir kelime dağarcığı ve kelimelerin morfolojik analizini kullanarak işleri düzgün bir şekilde yapmayı ifade eder (Manning, 2009). Bu tür tekniklerin kullanımı, algoritmaları eskisinden daha kolay hale getirdi.

Liu'ya (Bing, 2012) göre duygu analizi, insanların ürünler, hizmetler, organizasyonlar, bireyler, sorunlar, olaylar, konular ve bunların nitelikleri gibi varlıklara yönelik fikirlerini, duygularını, değerlendirmelerini, tutumlarını ve duygularını analiz eden çalışma alanıdır. Terim, Fikir Madenciliği ile birbirinin yerine kullanılabilir. Genel olarak, metne yönelik duygu, pozitif ve negatif kutupluluk olarak kategorize edilebilir. Duyarlılık analizi, analiz etme kapsamına göre belge düzeyi (tüm belge için duyarlılık hesaplanır), Cümle düzeyi (bir cümle için duyarlılık hesaplanır) ve Özellik/Varlık düzeyi (hedefin özelliklerini çıkar ve özellik akıllı polarite). Duygu analizini NLP teknikleriyle birleştirmek bize birçok avantaj sağlayabilir. Bu tür kombinasyonlar, tavsiye etme, ürün sıralaması, kamuoyu izleme, duygu tahmini gibi birçok alanda büyük başarılar elde etmiştir (Montoyo, 2012).

Duygu analizi yapmak için denetimli ve yarı denetimli öğrenme kullanılarak birçok yöntem geliştirilmiştir.

Denetimli öğrenme yönteminde, sınıflandırma modelleri standart duygu etiketli belgeler kullanılarak eğitilir. Denetimli sınıflandırmadaki en eski çalışmalardan biri film incelemeleri üzerine yapılmıştır. Veri seti, Naive Bayes, Maksimum Entropi Sınıflandırması ve Destek Vektör Makinesi (Bo, 2002) olmak üzere üç makine öğrenme modeli üzerinde eğitilmiştir. Bu araştırmaya dayanarak, derin öğrenme sistemlerinin geliştirilmesine yönelik çalışmalar genişletildi.

Dong ve Wei, 2014'te Twitter verilerini örnek olarak kullanarak duygu sınıflandırması için uyarlanabilir özyinelemeli Sinir Ağı (AdaRNN) önerdi (Dong, 2014). Derin öğrenme tekniklerindeki artışla birlikte, Twitter verileri üzerinde duygu analizi yapmak için Coooolll adlı bir derin öğrenme sistemi geliştirildi. Duyguları olumlu ve olumsuz kategorilere ayırmak için %87,61 doğruluk hesapladılar (Tang, 2014).

Yarı denetimli öğrenme yoluyla, denetimli yöntemler ve sözlük tabanlı yöntemler birleştirilebilir. Bu süreçte, sınıflandırma modelini eğitmek için az miktarda etiketli veri kullanılır ve büyük miktarda etiketlenmemiş veri alınır. Zhou ve Chen gibi birçok araştırmacı bunun üzerinde çalıştı ve hem eski hem de yeni etiketlenmemiş verileri kullanan bir yöntem önerdi. Duygu sınıflandırması için Bulanık derin inanç ağları adı verilen iki aşamalı yarı denetimli bir öğrenme yöntemi önerdiler. Model, film incelemeleri ve DVD incelemeleri veri setinde test edildi ve %79.5'lik bir doğruluk elde etti (Zhou, 2014).

Sözlük tabanlı yaklaşım, kelime veya cümlede metin duyarlılığı polarite sınıflandırması kavramına dayanmaktadır. Makine öğrenimi yaklaşımına benzer şekilde, birçok kişi bu yöntem üzerinde araştırma yapmıştır. Bir varsayıma göre, metnin duygu polaritesi, her bir kelimenin veya ifadenin duyarlılığının toplamıdır (Palanisamy, 2013). Aşağıdaki tablo, duygu hesaplanırken var olabilecek farklı sözlük türleri hakkında bir bakış sağlar.

Cümledeki her kelime farklı sözlük kategorilerine girdiğinden, duygu buna göre hesaplanabilir. Bir cümlenin duygu puanı, cümlenin her bir kelimesinin kutupluluklarının toplamını oluşturur.

Duygu tanıma, duygu analizinin derinliklerine giren bir çalışma alanıdır. En üst düzey duyarlılık analizi, yalnızca ifadeyi olumlu, olumsuz ve tarafsız etiketler olarak sınıflandırmaya yönelik olabilir. Ancak duygu, cümlelerin veya konuşmaların hangi yönden olumlu veya olumsuz olduğunu bilerek daha derine iner. İnsani duygu, yüz ifadeleri, yazılar, konuşmalar, eylemler ve jestler şeklinde elde edilebilir. Duygular alanında yapılan çalışmalar, araştırmacıları her zaman büyülemiştir ve birçok alana yayılmıştır.

Duygu analizi ile ilgili en alakalı çalışmalardan biri Alm ve ark. (Alm, 2005) metindeki duyguları otomatik olarak sınıflandırma görevini araştırmışlardır. Ekman, (Ekman, 1999) tarafından tanımlanan temel duygulara göre çocuk masallarındaki duyguları tanımlamışlardır. Yazarlar ayrıca cümleleri duygusal ve duygu olmayan olarak ve ayrıca olumlu duygu, olumsuz duygu ve duygu yok olan değerler olarak sınıflandırmışlardır. Ekman temel duygular kavramını kullanarak, tüm duygu sınıfları mutlu, üzgün, kızgın, tiksinti, korkmuş, olumlu şaşırmış ve olumsuz şaşırmış olarak ele almıştır. Bunlar arasında mutlu ve olumlu şaşkınlık olumlu duygu sınıfı, üzgün, kızgın, iğrenmiş, korkmuş ve olumsuz şaşırmış olanlar ise olumsuz duygu sınıfında sınıflandırılmıştır.

Duygu analizi çalışmalarında istatistiksel yaklaşımlar da kullanılmaktadır. İstatistiksel yaklaşımlar, hem Destek Vektör Makineleri (SVM) gibi geleneksel makine öğrenme tekniklerini hem de derin öğrenme modellerini kapsar. Bu yaklaşım, birkaç duygu analizi görevinde rekabetçi sonuçlar elde eden derin öğrenme modellerinin popülaritesi ve başarısı nedeniyle son yıllarda yükselişe geçmiştir (Rosenthal, 2017).

Derin Öğrenme tabanlı bir çalışmada ise Kim'in bir CNN kullanımını önerdiği ve %81.5'lik bir sınıflandırma doğruluğu elde ettiği gözlenmiştir (Kim, 2014). Socher arkadaşları tarafından yapılan diğer bir çalışmada bir cümle ve cümle düzeyindeki duygular için açıklamalı bir film inceleme veri seti kullanılmıştır (Socher, 2013). Wang ve arkadaşları tweet'lerde ikili polarite tahmini için bir Uzun-Kısa Süreli Bellek (LSTM) sinir ağı tanıtarak %87,2'lik bir doğruluk elde etmişlerdir (Wang, 2015). Bir başka çalışmada ise derin öğrenme tabanlı BERT ile yapılmıştır. Yin ve arkadaşları tarafından yapılan bu çalışmada ifade düzeyinde duygu sınıflandırması için BERT dil modeli kullanılarak eğitim işlemi gerçekleştirilmiştir (Yin, 2020).

Duygu analizi ile ilgili yapılan çalışmalardan bazıları ise hibrit tabanlı yaklaşımlar içermektedir. Hibrit tabanlı yaklaşımların temel amacının saf söz dizimleri tabanlı teknikler yerine anlam bilimsel ve farkındalık çerçevelerinin daha fazla dahil edilmesi olarak belirtilmiştir (Cambria, 2016). Yine bu kapsamda Gievska ve arkadaşları, insanların bir mobil uygulama aracılığıyla olumsuz duygularla baş etmelerine yardımcı olacak bir model geliştirmiştir. Bu hibrit model, duygulanımsal sözcükler içeren bir dizi sözcük kaynağı üzerine kurulmuştur ve son sınıflandırma için bir SVM algoritması kullanılmıştır (Gievska, 2014). Recupero ve arkadaşları, NLP ve Semantik Web teknolojilerini birleştiren denetimsiz, etki alanından bağımsız bir hibrit sistem olan 'Sentilo'yu daha da geliştirmeye odaklanmaktadır. Sentilo, ifade edilen bir görüşten bir dizi konuyu ve alt konuyu sınıflandırır ve ardından duyguyu değerlendirir. Bu model hem bir sözlük hem de fikir sahipleri arasındaki kavramları ve ilişkileri tanımlamak için kullanılan bir ontoloji kullanır (Recupero, 2015).

Yu ve arkadaşları, önceden eğitilmiş kelime yerleştirmelerine (örneğin: Word2Vec ve GloVe) uygulanabilecek bir vektör iyileştirme modeli önerdi; burada mevcut kelime yerleştirmeleri, anlamsal ve duygusal olarak benzer kelimeler birbirine daha yakın olacak ve bunun tersi olacak şekilde ayarlandı. Bu, gerçek değerli duygu puanlarını içeren bir duygu sözlüğü kullanılarak yapılır (Yu, 2017).

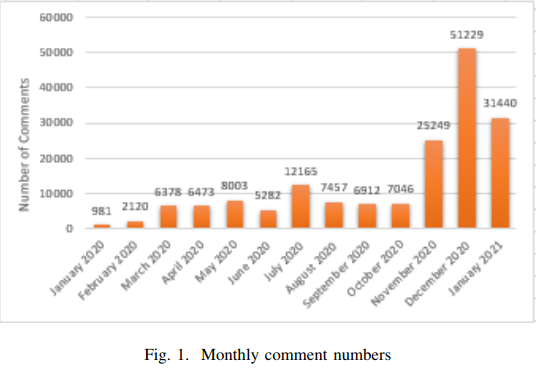


# MATERYAL ve YÖNTEM

# Materyal

# Veri Toplama

Sosyal medya kullanımı gün geçtikçe insanlar arasında daha önemli hale gelmektedir. Sosyal medyada insanların paylaşımları, yorumları vs. tüm etkileşimler birçok hayat problemine odaklanmak anlamında büyük önem arz etmektedir. Milyonlarca kullanıcısı olan sosyal medya platformlarındaki verilere erişmek artık daha zor hale gelmiştir. Twitter, Facebook, Instagram gibi sosyal medya platformları kullanıcılarının kişisel haklarını korumak adına verilerin çekilmesi anlamında çok ciddi kısıtlamalar uygulamaktadır. Reddit sosyal medya platformu ise verilerin çekilmesi anlamında daha fazla kolaylık sağlamaktadır.

 şekli değiştir…

* + - 1. **Reddit**

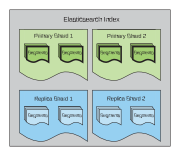
Web verileri, özellikle uygulama programlama arayüzlerinden (API'ler) gelen veriler, çevrimiçi sosyal platformların kullanıcı tarafından oluşturulan etkinlik ve içerik veritabanlarını kullanan araştırmacılar için muazzam bir nimet olmuştur [1] . Pushshift Reddit veri kümesi ayrıca araştırmacı erişimi için bir API ve araştırmacıların toplanan verilerle kolayca etkileşim kurmasını sağlayan bir Slackbot içerir. Pushshift Reddit API, araştırmacıların aylık dökümleri indirmeye gerek kalmadan tüm veri setinde kolayca sorgu yürütmesini sağlar.

Pushshift, verileri toplamak, depolamak, kataloglamak, indekslemek ve son kullanıcılara dağıtmak için birden çok arka uç yazılım bileşeni kullanır. Şekil 1'de görüldüğü gibi, bu alt sistemler şunlardır:

Ham verilerin toplanmasından ve depolanmasından sorumlu olan alma motoru.

Gelişmiş veri sorgulamasına ve meta-veri depolamasına izin veren bir PostgreSQL veritabanı.

Alınan verilerin indekslenmesini ve toplanmasını gerçekleştiren bir Elastic Search belge deposu kümesi



Alma Motoru

Sorgulanabilir Veri

API

PostgreSQL

Arşiv

Şekil 1: Reddit Pushshift Veri Toplama Platformu.

* + 1. **Verilerin Çekilmesi**

Verileri çekmek için DataExtract adında aylara göre belirlenen alt konu başlıklarında verileri çekmemize olanak sağlayan python programlama dilinde bir sınıf tanımlaması yaptık. Konu başlıkları, ay sayısı gibi verileri dışarıdan girmemize olanak sağlayan bu sınıf sayesinde farklı alanlarda da verilerin çok daha kolay bir şekilde çekilmesine olanak sağlandı. Ayrıca veriler çekilirken kullanıcılara ait yorumları ayrı bir json dosyasında yedekledik. Bu şekilde kullanıcıların da tüm yorumlarının genel bir duygu sınıflandırmalarına ilişkin bilgi edinildi. Kullanıcı yorumlarından ortalama duygular çıkarılıp kullanıcıların aylara göre duygu değişimlerini görme fırsatı da elde edilmiş oldu.



Grafiği değiştir buraya ekle.

Tablo 1 de PushShift API ile çekilen verilere ait bazı alanlar ve açıklamaları verilmiştir. Özellikle selftext alanındaki yorumlar üzerinde işlemler gerçekleştirilmiştir.

|  |  |
| --- | --- |
| **Alan** | **Açıklama** |
| **id** | The gönderiyi yapan, örneğin., “3lhg2” |
| **url** | Gönderinin gönderildiği internet adresi bilgisi.  Örneğin, [“https://www.reddit.com/r/Coronavirus/](http://www.reddit.com/r/AskReddit/) |
| **permalink** | Gönderiye ait kalıcı bağlantı adresi |
| **author** | Gönderiyi yapanın kullanıcı adı |
| **created utc** | Gönderinin oluşturulma zamanı |
| **subreddit** | Gönderinin gönderildiği konu başlığı |
| **subreddit id** | Gönderinin gönderildiği konu başlığına ait id |
| **selftext** | Gönderiye ait içeriğin olduğu alan |
| **title** | Gönderi başlığı |
| **num comments** | Gönderiye yapılan yorum sayısı |
| **score** | Gönderiye yapılan yorumlar ve beğeniler sonucu verilen puan |
| **is self** | Gönderinin kişiye ait olup olmadığını belirten kontrol durumu |
| **over 18** | Gönderinin güvenli olup olmadığını belirten kısım |

Tablo 1. PushShift’den çekilen verilere ait alanlara ait bilgiler

* + 1. **Verilerin Temizlenmesi**

Sosyal Ağ platformları, kullanıcıların gönderiler ve yorumlar gibi potansiyel olarak sonsuz miktarda metin içeriği oluşturmasına olanak tanır. Bu platformlarda kullanılan dil genellikle yapılandırılmamış veya yarı yapılandırılmıştır. Kullanıcılar, bir cümle kurarken, yazım veya dilbilgisi kurallarını dikkate almama eğilimindedir, genellikle kendilerini konuşma dili veya diyalektik ifadelerle ifade ederler. Bu sözdizimsel ve anlamsal belirsizlikler, yapılan yorumlar üzerinde gerçekleştirmek istediğimiz sınıflandırma işlemleri için olumsuzluklar oluşturacaktır.

NLP ve Metin Madenciliği araştırma alanları, bu sorunları azaltmak için teknikler ve metodolojiler sunar. NLP ifadesi, doğal dilin bilgisayar ortamında işlenmesini amaçlayan teknikleri tanımlar. Metin madenciliği terminolojisi, metin içeriğinden otomatik olarak bilgi çıkarmak için teknolojilerin kullanılması anlamına gelir.

Bu çalışmada, ilk adım metin ön işleme ile ilgilidir. Bu kavram, hem anormallikleri tespit etmeyi ve kaldırmayı hem de analiz edilen metnin boyutunu küçültmeyi amaçlayan yapılandırılmamış metinsel verilerin tüm ön işleme ve temizleme işlemlerini ifade eder. Geleneksel metin ön işleme ardışık düzeni aşağıdaki işlemleri içerir:

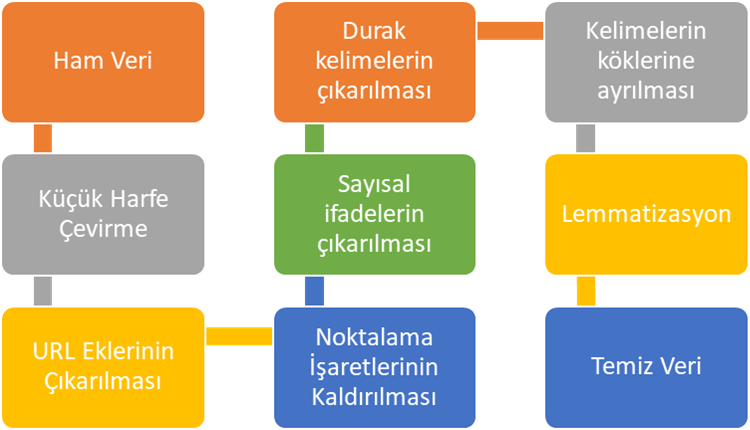
**Durdurulan sözcüklerin ortadan kaldırılması:** Durdurulan sözcükler, bir dilde sıklıkları göz önüne alındığında, ilgili bilgilerin çıkarılması amacıyla genellikle çok az öneme sahip olduğu düşünülen sözcüksel olarak boş sözcüklerdir (örneğin, İngilizce'de the, a, be). Bu kelimelerin kaldırılması, metnin boyutunu önemli ölçüde küçülttüğü için, metnin ön işlemesinin verimliliğini ve etkililiğini artırır.

**Küçük harf:** Bu terminoloji, bir metindeki tüm kelimelerin küçük harfe dönüştürülmesini ifade eder. Esas olarak girdinin tutarlılığı ve uyumu nedenleriyle benimsenmiştir.

**Simgeleştirme:** Bu ifade, bir metnin, simge adı verilen, onu oluşturan sözcüklere bölünmesi sürecini tanımlar. Genellikle noktalama işaretleri bu aşamada kaldırılır.

**Lemmatization:** lemmatization, bir kelimenin çekimli formunun lemma adı verilen kanonik formuna indirgeme sürecini ifade eder. Bu aşama, çıktının seyrekliğini azaltmayı amaçlar.

Veri Çekilmesi kısmında belirttiğimiz DataExtract isimli sınıf içerisinde yukarıda tanımlanan veri temizleme işlemlerini dataProcessing isimli fonksiyon ile tanımladık. Böylece veriler çekilirken aynı zamanda temizleme işlemi de gerçekleştirilmiştir.



Şekil 2. Veri Temizleme Aşamaları

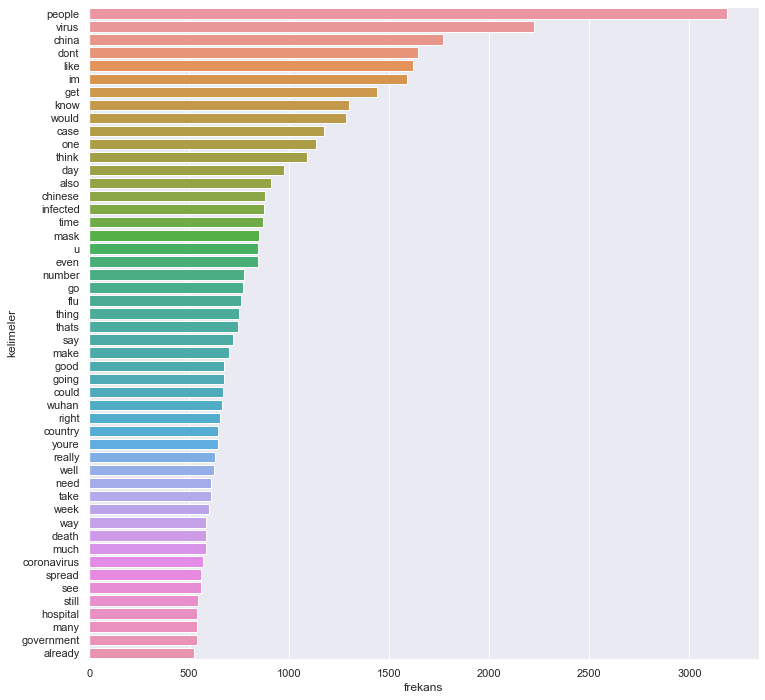
Konu başlıkları belirlenirken ise çalışmanın ana konusu olan koronavirüs salgını ve koronavirüs aşısı ile ilgili en çok etkileşim içeren konu başlıklarına odaklandık. antivaxx, DebateVaccines, vaxxhappened, Vaccines, CovidVaccinated gibi en çok etkileşim alan konu başlıklarından veriler çekilmiştir.

Veriler çekilip temizlendikten sonra üzerinde hızlıca sınıflandırma ve eğitim işlemleri yapabilmek için clean\_data.db isminde bir veritabanında veriler tablo olarak tutulmuştur. Bütün bu işlemler sonrasında ise artık verilerimiz sınıflandırma işlemleri için hazır hale gelmiştir.

Tablo 2 de temizlenmiş verilere ilişkin farklı tarihlerde farklı subreddit içerisinden bazı örnekler verilmiştir. Çalışmalarımız İngilizce veriler üzerinde etkili olduğu için ingilizce örnekler seçilmiştir.

|  |  |  |  |
| --- | --- | --- | --- |
| **subreddit** | **id** | **date** | **clean\_text** |
| vaxxhappened | fco8iq6 | 01/01/2020 | you think that after seeing him suffer tetanus theyd at least give him the tetanus shot |
| Vaccines | fehpvnq | 15/01/2020 | thank you i know its the best choice for me baby and others around me |
| CovidVaccinated | hqh7srg | 29/12/2020 | im in the states in a region where none of those restrictions apply i want to be protected but if pfizer shots only get me to protected like wtf |
| antivaxx | gho2jvl | 01/01/2021 | every vaccine has destroyed portions of my brain to a point i can no longer give a damn about anything people say its just my attitude but i know better its vaccines eating away at my brain i reckon i will wake up a zombie soonoh damn i just realized this is how it all begins |
| DebateVaccines | frvf51n | 26/05/2020 | example see postmarketing experience because these events are reported voluntarily from a population of uncertain size it is not always possible to reliably estimate their frequency or establish a causal relationship to vaccination |

Tablo 2. Temizlenmiş verilere ait bazı örnekler



* 1. **Yöntem**

Bu bölümde reddit platformunda ilgili subreddit’lerden çekilen yorumları positive, negative ve neutral olarak 3 sınıfta etiketlemeye çalıştık. Etiketleme işlemi için VADER (Gilbert, 2014) sınıflandırıcısı kullanılmıştır.

* + 1. **Duygu Polarite Sınıflandırması**

Duygu polarite sınıflandırması, duygu analizi alanında en çok çalışılan problemdir. Tipik olarak, görev çok sınıflı bir sınıflandırma problemi olarak kabul edilir: Subjektif bir metin verildiğinde amaç, metnin genel tonunun olumlu, olumsuz veya nötr olup olmadığını belirlemektir. Bu görev, çeşitli ayrıntı düzeylerinde gerçekleştirilebilir. Sözcüklerin ve ifadelerin duygu kutupluluğu ilişkilendirmelerinden, cümlelerin, SMS mesajlarının, sohbet mesajlarının ve tweet'lerin duyarlılığına, ürün incelemelerinde, blog gönderilerinde ve bütünlükteki duygu analizine kadar belgeler.

**Kelime Düzeyi:** Bu analiz seviyesinin fikri, bir kelimeye bir duygu değeri atamaktır. Girişler adı verilen bir sözcük-duygu ilişkileri topluluğu, bir duygu sözlüğü oluşturur. Her giriş, kelime olan bir anahtar ve pozitif veya negatif olabilen bir değer veya kelime ile duygu kategorisi arasındaki ilişkinin gücünü gösteren gerçek bir değerden oluşur. Bu tür sözlükler ya manuel açıklama ile ya da otomatik olarak oluşturulabilir (Kiritchenko, 2014). El ile oluşturulan sözlükler genellikle birkaç bin girdi içerir. Buna karşılık, otomatik olarak oluşturulan sözlükler, yüz binlerce kelime veya kelime dizisi için duygu ilişkilerini yakalayabilir.

**Cümle Düzeyi:** Cümle düzeyindeki duygu analiz sistemleri, tüm cümlelere olumlu, olumsuz veya nötr gibi etiketler atamayı amaçlar. Şunu belirtmekte fayda var ki, bir cümlenin duygu düzeyi, sadece kelimelerinin kutuplarını bir araya getirerek elde edilemez. Bu nedenle, bağlamı anlayan ve cümlelerde ifade edilen doğru duyguyu çıkaran Makine Öğrenimi tabanlı sistemler geliştirmek gerekir. Bu tür sistemler, metinden ve/veya dış kaynaklardan, örneğin duygu sözlüklerinden çıkarılan çok sayıda özelliği kullanarak etiketlenmiş eğitim verilerinden (zaten olumlu, olumsuz veya nötr olarak işaretlenmiş cümleler) bir eşleme öğrenir.

**Belge Düzeyi:** Belge düzeyinde duygu analizi genellikle bileşen cümlelerin duygu analizine ayrıştırılır. Ancak, duyguyu tüm belgelerde özetlemeyi öneren bazı çalışmalar vardır (Liu, 2011).

* + 1. **VADER**

VADER (Valence Aware Dictionary ve sEntimentReasoner), sözlük ve kural tabanlı bir duygu analizi aracıdır. George Berry, Ewan Klein ve Pier Paolo tarafından geliştirilen MIT lisansı altında bir açık kaynaktır. Vader sözlüğü, sosyal medya alanında son derece iyi bir performans sergiliyor. VADER, LIWC (Linguistic Inquiry ve Word Count) gibi geleneksel duygu sözlüklerinin faydalarını korur. Daha büyüktür, basitçe incelenir, anlaşılır, hızlı uygulanır ve kolayca genişletilir. VADER duygu sözlüğü altın standart kalitedir ve insanlar tarafından onaylanmıştır.

Genellikle anlamsal yönelimlerine göre ya olumlu ya da olumsuz olarak etiketlenen sözcüksel özelliklerin bir listesi olan duygu sözlüğünün bir kombinasyonunu kullanır. VADER, sosyal medya metinleri, film incelemeleri ve ürün incelemeleriyle uğraşırken oldukça başarılı olmuştur. Bunun nedeni, VADER'in yalnızca olumluluk ve olumsuzluk puanını değil, aynı zamanda bir duygunun ne kadar olumlu veya olumsuz olduğunu da söylemesidir. VADER geliştiricileri, derecelendirmelerinin çoğunu almak için Amazon'un Mechanical Turk'ünü kullandı.

VADER, metindeki kelimelerden herhangi birinin VADER sözlüğünde olup olmadığını görmek için bir metin parçasını analiz eder. polarity\_scores() fonksiyonunu kullanarak polarite indekslerini bulabilir. Bu, belirli bir cümle için negatif, nötr, pozitif ve bileşiğin metrik değerlerini döndürür. Bileşik puan, -1 ile +1 arasında normalleştirilmiş tüm sözlük derecelendirmelerinin toplamını hesaplayan bir ölçümdür; burada -1 en aşırı olumsuzu ve +1 en aşırı olumluyu belirtir. Cümleleri olumlu, tarafsız veya olumsuz olarak sınıflandırmak için standartlaştırılmış eşikler belirlemek yararlıdır. Tipik eşik değerleri aşağıda verilmiştir.

Olumlu Duygu: bileşik puan >= 0.05

Tarafsız Duygu: bileşik puan > -0.05 ve <0.05

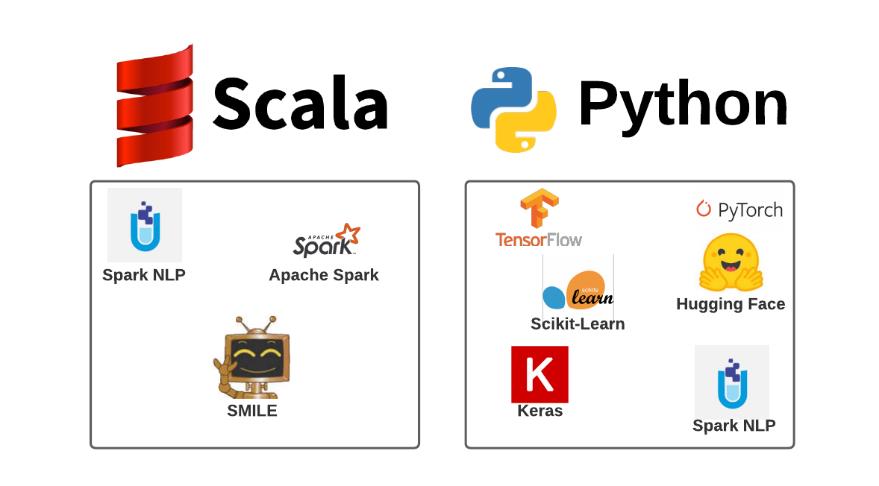
Negatif Duygu: bileşik puan <= -0.05

Bunlar, belirli bir metin incelemesi için çok boyutlu duyarlılık ölçümleri için en yararlı ölçümlerdir. Aşağıdaki şekil, sözcükleri ve duygu derecelendirmelerini içeren VADER sözlüğünü göstermektedir.

VADER, duyguları öncelikle Noktalama İşaretleri, Büyük Harfler, Derece değiştiriciler, Bağlaçlar, Önceki Tri-gram gibi belirli kilit noktalara dayalı olarak analiz eder.

Hem duygu polaritesini hem de -4 ila +4 arasında değişen duygu yoğunluğunu gösteren doğrulanmış değerlik puanlarına sahip 7.500'den fazla sözcüksel özellik vardır.

* + 1. **sparkNLP**

****

Spark NLP, Apache Spark ve SparkML [13] üzerine inşa edilmiş açık kaynaklı bir metin işleme kitaplığıdır. Doğal Dil İşleme ardışık düzenleriyle kullanılabilecek bir API sağlar. Kitaplık, özel modeller oluşturmanın yanı sıra eğitimli modeller ve işlem hatları kullanmak için destek sağlar.

Çerçeve, farklı dil temsil modelleri sağlar. Desteklenen modeller, çeşitli metin sınıflandırma görevlerini gerçekleştirmek için yeniden kullanılabilir. Spark NLP, iki tahminci kullanarak duyarlılık analizinin uygulanmasını mümkün kılar: SentimentDLApproach ve ClassifierDLApproach.

SentimentDLApproach, pozitif veya negatif olarak etiketlenmiş metni kullanarak sınıflandırmayı öğrenir. Olumlu veya olumsuz olarak sınıflandırılıp sınıflandırılmayacağı çok büyükse, metni nötr olarak sınıflandırır. Kesinlik düzeyi, sınıflandırıcı tarafından kabul edilen bir sınıflandırmaya ait metnin olasılığı olarak tanımlanır. Belirsizlik, ayarlanabilir bir eşiğin altındaki bir olasılıktır.

ClassifierDLApproach, etiketli metin içeren bir veri kümesini kullanarak sınıflandırmayı öğrenir. Veri kümesi tarafından kullanılan tüm benzersiz etiketler olası sonuçları temsil edecektir. İkili duyarlılık analizi yapmak için pozitif veya negatif olarak etiketlenmiş metin kullanılacaktır. Çok sınıflı duygu analizi için pozitif, nötr veya negatif içeren metin kullanılmalıdır.

* + 1. **BERT**

Transformers'dan Çift Yönlü Kodlayıcı Temsilleri (BERT), Google AI ekibi tarafından 2018'de derin öğrenme tekniklerine dayalı olarak önerilen önceden eğitilmiş bir dil temsil modelidir (Devlin ve diğerleri, 2019). Diğer dil temsil modellerinden farklı olarak, tüm katmanlarda hem sol hem de sağ bağlamda ortak koşul aracılığıyla BERT, etiketlenmemiş giriş metninden derin çift yönlü temsiller üretebilir. BERT, metin sınıflandırma ve soru cevaplama gibi çeşitli NLP görevlerinde uygulanmış ve mükemmel bir performans sergilemiştir (Yuwen ve Zhaozhuo, 2018).

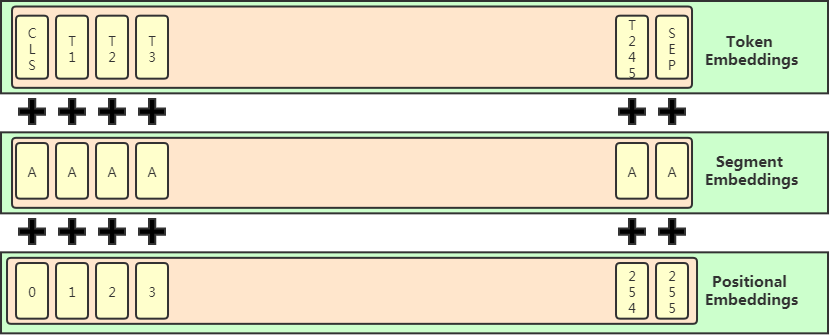
Benimsenen ince ayar yaklaşımı nedeniyle, BERT kullandığımızda aşağı akış NLP görevleri için belirli bir mimari yoktur. Akıllı bir etmen olarak, model tasarımında önceki insan bilgisinin kullanımını en aza indirmeli ve bu bilgileri veri yerine veriden öğrenmelidir. BERT'de, bir sonraki kelime tahmininin sık kullanılan amacından ziyade dil modelini eğitmek için kullanılan iki farklı hedef vardır: Birincisi, maskeli dil modeli hedefidir ve modelin maskelenmiş belirteçleri bağlamlarından tahmin etmesi gerekir. Diğeri, modelin B dizisinin A dizisini takip edip etmediğini öğrenmesi gereken bir sonraki dizi tahmin hedefidir. Bu iki amaç, modelin uzun vadeli bağımlılıkları daha iyi öğrenmesini sağlar.

Maskeli Dil Modeli Amaç: Model, A dizisi ve B dizisinde rastgele maskelenen belirteçleri tahmin etmeyi öğrenir.

Sonraki Cümle Tahmini: BERT'nin uzun vadeli bağımlılıkları daha iyi öğrenmesini sağlamak için, modelin bir B dizisinin doğal olarak önceki A dizisini izleyip izlemeyeceğini öğrenmesi gerekir. Dolayısıyla A dizisi ve B dizisi aynı belgedendir, böylece A dizisi B dizisini takip eder.

BERT'de (Devlin ve diğerleri, 2019), yazarlar transformatörü tekrarlayan veya evrişimli sinir ağları yerine temel bileşenler olarak kullanırlar. Transformatör yalnızca kendi kendine dikkat mekanizmasına dayanmaktadır. Tekrarlayan Sinir Ağı (RNN) veya Evrişimli Sinir Ağı (CNN) ile karşılaştırıldığında, transformatörün üç avantajı vardır. İlk olarak, hesaplama kaynağını ve hesaplama hızını azaltabilir. İkinci olarak, RNN'de imkansız olan hesaplama paralelleştirilebilir. Aksi takdirde, transformatör uzun menzilli bağımlılıkları öğrenmede iyi bir performansa sahiptir.

Pratikte, metin sınıflandırma ve soru yanıtlama gibi çeşitli NLP görevleri için çok fazla önemli göreve özgü mimari değişiklik yapmadan BERT'ye bir ek çıktı katmanıyla ince ayar yaparak mükemmel bir performans modeli oluşturmak kolaydır (Devlin ve diğerleri, 2019).



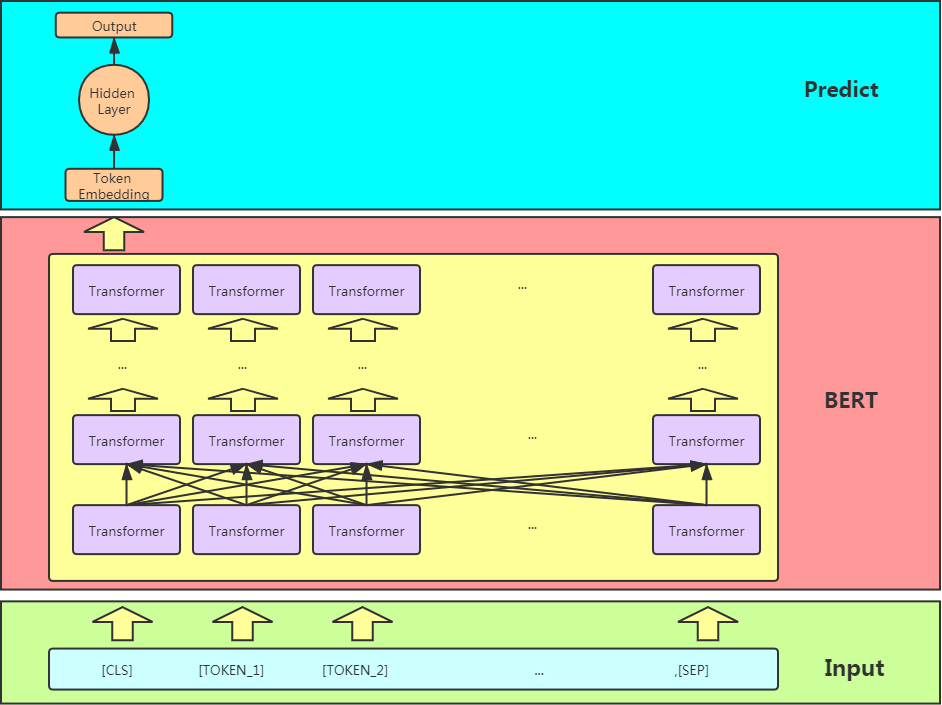
**Figure 10**. The input representation (Devlin et al., 2019).

Google, Github'da önceden eğitilmiş BERT'yi farklı boyutlarda cilaladı. Deneylerimiz için bir temel olarak, Google araştırma ekibi tarafından önceden eğitilmiş, 12 katman ve belirteç başına 768 gizli boyut ile BERT tabanı kullanılmıştır. Toplam 110 milyon parametreye sahiptir. Modelin mimarisi Şekil 11'de aşağıdaki üç parça ile sunulmaktadır.

Giriş: İlk kısım giriş katmanıdır. Sorumluluğu, ilk gömme kelimesini kabul etmek ve onu BERT'e teslim etmektir. İlk sözcük yerleştirmenin nasıl yapılacağı Bölüm 3.2'de açıklanmıştır.

BERT: 𝐵𝐸𝑅𝑇(·) önceden eğitilmiş BERT modeli olsun, x giriş kısmından ilk gömme olsun. Önce gizli gösterimi 𝑍 = 𝐵𝐸𝑅𝑇(𝑥) ∈ 𝑅𝑟ℎ∗|𝑥| olarak elde ederiz, burada |𝑥| giriş dizisinin uzunluğu ve 𝑟ℎ gizli boyutun boyutudur. Bu bölümün çıktısı, her bir girdi belirtecinin son sözcük gömmesidir.

Tahmin: Gizli gösterim yoğun bir katmana ve ardından 𝑠𝑜𝑓𝑡𝑚𝑎𝑥 fonksiyonlarına iletilir: 𝑔 = 𝑠𝑜𝑓𝑡𝑚𝑎𝑥(𝑤 · 𝑍 + 𝑏) = 𝑠𝑜𝑓𝑡𝑚𝑎𝑥(𝐵𝐸𝑅𝑇(𝑥)) , burada 𝑊 ∈ 𝑅𝑟ℎ 𝑎𝑛𝑑 𝑏 ∈ 𝑅 . 𝑠𝑜𝑓𝑡𝑚𝑎𝑥, dizinin boyutu boyunca uygulanır. Matematikte, softmax, girdi olarak K reel sayı vektörünü alan ve onu girdi sayılarının üstelleriyle orantılı K olasılıklarından oluşan bir olasılık dağılımına normalleştiren bir fonksiyondur. Çıktı, her bir etiketin olasılıklarıdır: 𝑝𝑟𝑒𝑑𝑖𝑐𝑡\_𝑙𝑎𝑏𝑒𝑙 = 𝑎𝑟𝑔 𝑚𝑎𝑥(𝑔), burada çıktı ∈ R𝑘 ve 𝑘, veri kümesinde pozitif, negatif ve nötr dahil olmak üzere üç duygu olduğu için 3'e eşittir.



**Figure 11**. The architecture of BERT used in sentiment classification.

Modeli eğitirken, ağırlık kayıp fonksiyonu ile ayarlanır ve kayıp fonksiyonunu en aza indirerek uygun bir BERT modeli oluşturulur. İleri yayılım sırasında, modelin çıktıları olası etiketlerin olasılıklarıdır. Bu olasılıklar hedef etiketlerle karşılaştırılır. Ardından, kayıp fonksiyonu, hedef etiket ile modelin çıktıları arasındaki herhangi bir sapma için bir ceza hesaplar. Geri yayılım sırasında, eğitilebilir ağırlıklar, her ağırlık için kayıp fonksiyonunun kısmi türevi hesaplanarak ayarlanır. Normal şartlar altında, veri seti üzerinde eğitimden sonra daha düşük kayıplı bir model üretilir.

* + 1. **Değerlendirme Metrikleri**

Derin öğrenme modelleri yapıcı bir geri bildirim ilkesine göre çalışmaktadır. Öncelikle bir model oluşturulur, ölçümlerden geri bildirim alınır, iyileştirmeler yapılır ve arzu edilen bir doğruluğa ulaşana kadar devam edilir. Deneyler farklı modellerle gerçekleştirilir ve her bir deneyin sonucu bir metrik ile ölçülür. Bu aşamalarda değerlendirme ölçütleri bir modelin performansını açıklamaktadır. Değerlendirme ölçütlerinin önemli bir yönü, model sonuçlarını ayırt etme yetenekleridir.

Bu çalışmada, melanom tespiti için eğitilen modelin test edilmesinde doğruluk(accuracy)(Denklem 3.1), kesinlik(precision)(Denklem 3.2), hatırlama(recall) (Denklem 3.3) ve F ölçümü(f1-score)(Denklem 3.4) gibi çeşitli standart değerlendirme ölçütleri kullanılmaktadır. Ayrıca testlerde elde edilen sonuçlar Şekil 3.14’da gösterilen karmaşıklık matrisi(confusion matrix) üzerinde verilmektedir. Denklemlerde, gerçek pozitifler (TP), doğru şekilde tahmin edilen örneklerin sayısıdır; yanlış negatifler (FN), yanlış tahmin edilen örneklerin sayısıdır. Doğru negatifler (TN), doğru şekilde tahmin edilen negatif örneklerin sayısıdır, yanlış pozitifler (FP) ise yanlış tahmin edilen negatif örneklerin sayısıdır.

(3.1)

(3.2)

(3.3)

(3.4)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | | Tahmin Edilen | |
|  |  | | Pozitif | Negatif |
| Gerçek Değer | | Pozitif | Doğru Pozitif (TP) | Yanlış Negatif (FN) |
| Negatif | Yanlış Pozitif (FP) | Doğru Negatif (TN) |

Şekil 3.14. Karmaşıklık Matrisi

3 Sınıflı Duygu Sınıflandırma ve Eğitim Sonuçları

Çizelge 4.2. Vader ile 3 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri

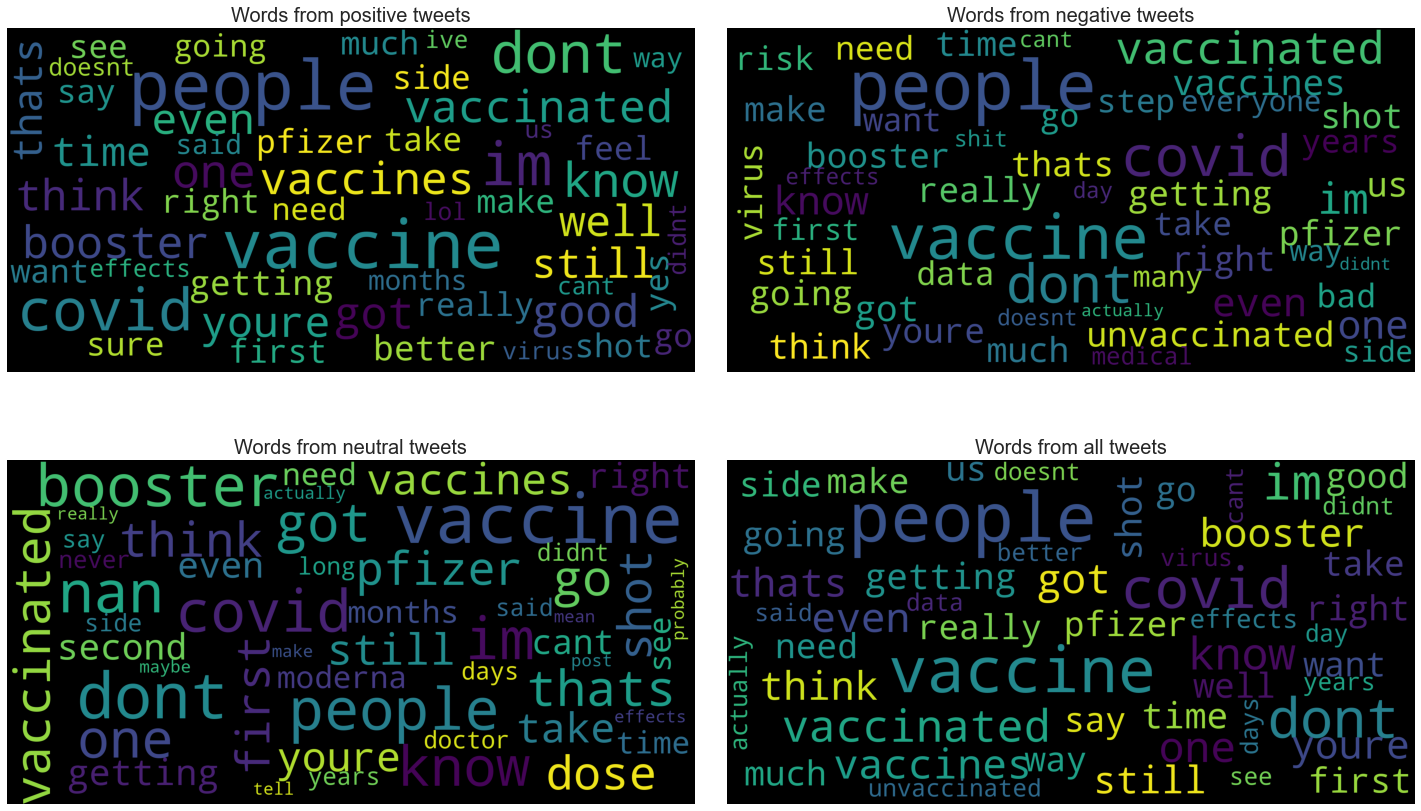
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sentiment  (Duygular) | Hassasiyet  (Precision) | Geri Çağırma (Recall) | F1-skoru  (F1-Score) | Doğruluk  (Accuracy) |
| Negative | 0.68 | 0.75 | 0.71 | 0.70 |
| Positive | 0.74 | 0.70 | 0.72 |
| Neutral | 0.69 | 0.64 | 0.67 |

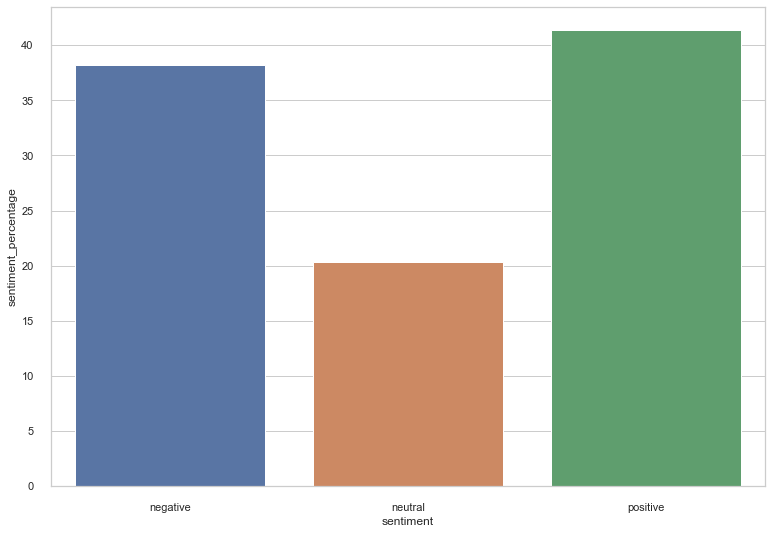
Çizelge 4.2. Glove ile 3 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sentiment  (Duygular) | Hassasiyet  (Precision) | Geri Çağırma (Recall) | F1-skoru  (F1-Score) | Doğruluk  (Accuracy) |
| Negative | 0.72 | 0.79 | 0.75 | 0.75 |
| Positive | 0.76 | 0.74 | 0.76 |
| Neutral | 0.73 | 0.68 | 0.72 |

Çizelge 4.2. Bert ile 3 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sentiment  (Duygular) | Hassasiyet  (Precision) | Geri Çağırma (Recall) | F1-skoru  (F1-Score) | Doğruluk  (Accuracy) |
| Negative | 0.80 | 0.86 | 0.83 | 0.82 |
| Positive | 0.83 | 0.79 | 0.81 |
| Neutral | 0.82 | 0.78 | 0.80 |





4 Sınıflı Duygu Sınıflandırma ve Eğitim Sonuçları

Çizelge 4.2. sparkNLP ile 4 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri

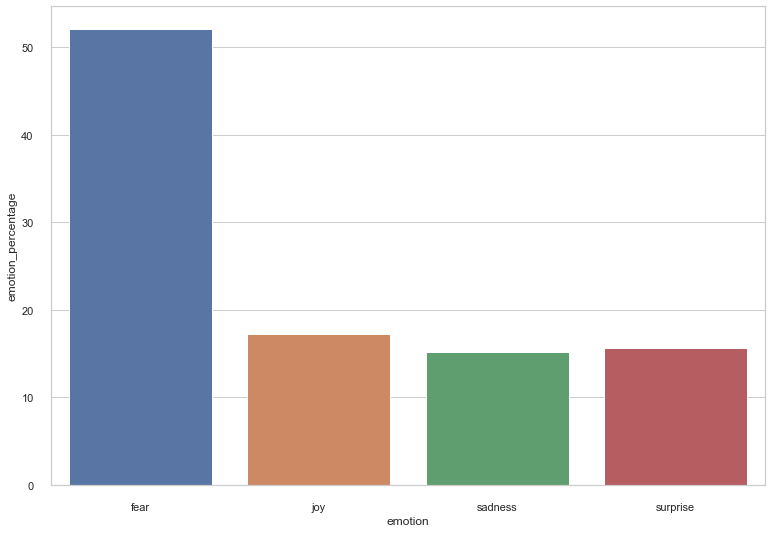
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Emotions  (Duygular) | Hassasiyet  (Precision) | Geri Çağırma (Recall) | F1-skoru  (F1-Score) | Doğruluk  (Accuracy) |
| Fear  (Korku) | 0.71 | 0.76 | 0.73 | 0.72 |
| Sadness  (Mutsuzluk) | 0.75 | 0.72 | 0.74 |
| Joy  (Eğlenceli) | 0.72 | 0.67 | 0.70 |
| Surprise  (Şaşırtıcı) | 0.73 | 0.71 | 0.72 |

Çizelge 4.2. Glove ile 4 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Emotions  (Duygular) | Hassasiyet  (Precision) | Geri Çağırma (Recall) | F1-skoru  (F1-Score) | Doğruluk  (Accuracy) |
| Fear  (Korku) | 0.73 | 0.77 | 0.74 | 0.75 |
| Sadness  (Mutsuzluk) | 0.76 | 0.73 | 0.75 |
| Joy  (Eğlenceli) | 0.73 | 0.69 | 0.72 |
| Surprise  (Şaşırtıcı) | 0.75 | 0.73 | 0.74 |

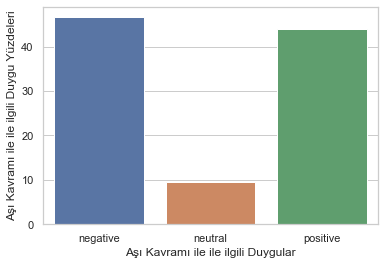
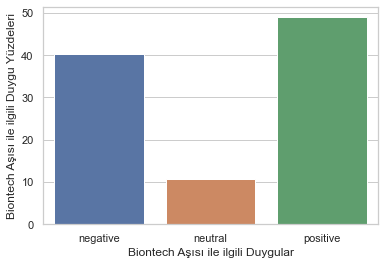
Çizelge 4.2. Bert ile 4 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri

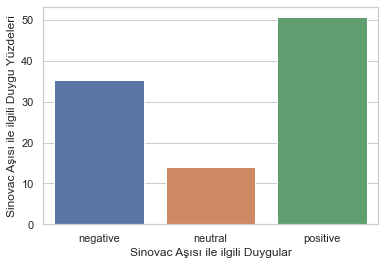
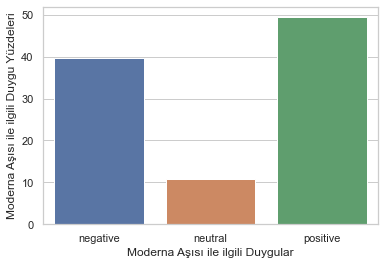
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Emotions  (Duygular) | Hassasiyet  (Precision) | Geri Çağırma (Recall) | F1-skoru  (F1-Score) | Doğruluk  (Accuracy) |
| Fear  (Korku) | 0.76 | 0.78 | 0.76 | 0.78 |
| Sadness  (Mutsuzluk) | 0.77 | 0.75 | 0.78 |
| Joy  (Eğlenceli) | 0.75 | 0.72 | 0.74 |
| Surprise  (Şaşırtıcı) | 0.78 | 0.74 | 0.77 |

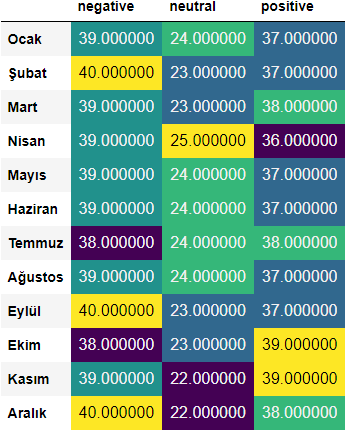
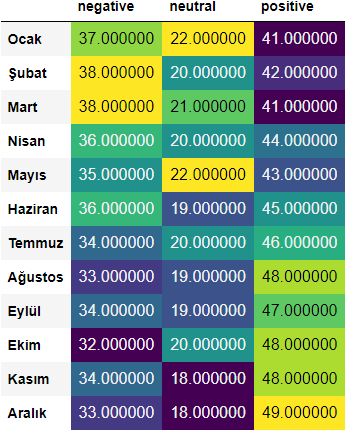


Aşı Analizi

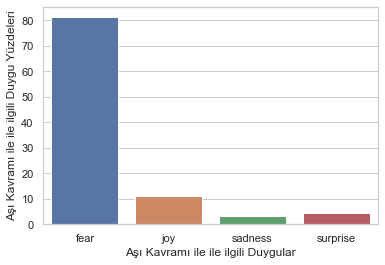
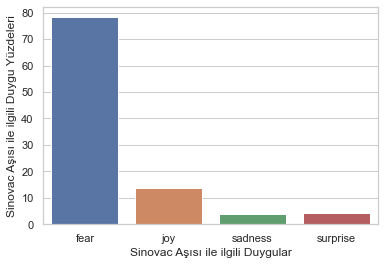
3 Sınıflı

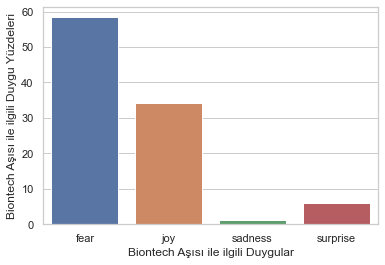
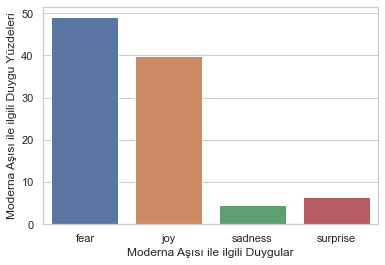
 

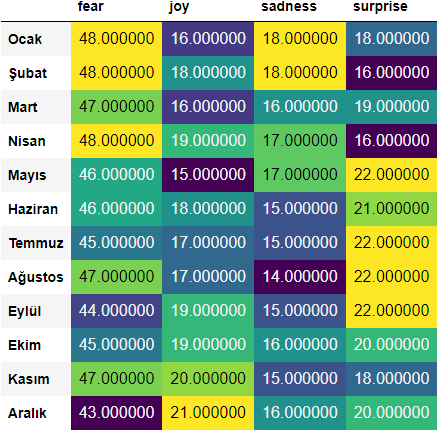
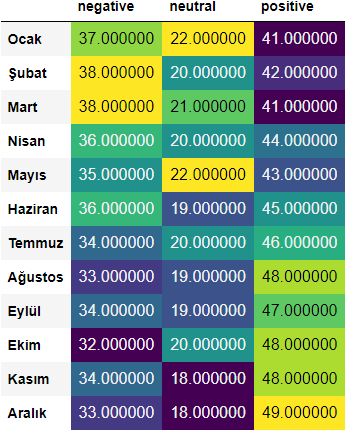
 

4 sınıflı

HAND, D.J., 2007. Principles of Data Mining. Drug-Safety 30, 621–622.

BAUMGARTNER, J.; ZANNETTOU, S.; KEEGAN, B.; SQUIRE, M.; BLACKBURN, J., 2020. The Pushshift Reddit Dataset. Proceedings of the International AAAI Conference on Web and Social Media, [S. l.], v. 14, n. 1, p. 830-839.

SINGHAL, T., 2020. A Review of Coronavirus Disease-2019 (COVID-19). Indian J Pediatr 87, 281–286.

WEIDMAN, A. C., STECKLER, C. M., & TRACY, J. L., 2016. The Jingle and Jangle of Emotion Assessment: Imprecise Measurement, Casual Scale Usage, and Conceptual Fuzziness in Emotion Research. Emotion, 17(2), 267–295.

CALVO, R. A. and KIM, S. M., 2010. Sentiment Analysis in Student Experiences of Learning. Third International Conference on Educational Data Mining (EDM2010), 111–120.

DENG, L., and DONG, Y., 2014. "Deep learning: methods and applications." Foundations and Trends® in Signal Processing 7.3–4: 197-387.

KHURANA, DIKSHA, KOLI, A., KHATTER, K. and SINGH, S., 2017. "Natural language processing: State of the art, current trends and challenges." arXiv preprint arXiv:1708.05148.

MANNING, C. D., RAGHAVAN, P. and SCHUTZE, H., 2009. “An Introduction to Information Retrieval” online, Retrieved from the internet.

BING, L., 2012. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5, no. 1:1-167.

MONTOYO, ANDRES, MARTINEZ-BARCO, P. and BALAHUR, A., 2012. "Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments." Decision Support Systems 53, no. 4:675-679. 

PANG, BO, LEE, L. and VAITHYANATHAN, A., 2002. "Thumbs up? Sentiment classification using machine learning techniques." arXiv preprint cs/0205070.

TANG, DUYU, WEI, F., QIN, B., LIU, T. and ZHOU, M., 2014. "Coooolll: A deep learning system for twitter sentiment classification." In Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), pp. 208-212.

ZHOU, SHUSEN, CHEN, Q. and WANG, X., 2014. "Fuzzy deep belief networks for semi-supervised sentiment classification." Neurocomputing 131: 312-322.

PALANISAMY, PRABU, YADAV, V. and ELCHURİ, H., 2013. "Serendio: Simple and Practical lexicon based approach to Sentiment Analysis." In Second Joint Conference on Lexical and Computational Semantics (\* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 543-548.

ALM, OVESDOTTER, C., ROTH, D. and SPROAT, R., 2005. "Emotions from text: machine learning for text-based emotion prediction." In Proceedings of human language technology conference and conference on empirical methods in natural language processing, pp. 579-586.

EKMAN, P., 1999. "Basic emotions." Handbook of cognition and emotion 98, no. 45-60.

ROSENTHAL, S., FARRA, N. and NAKOV, P., 2017. Semeval-2017 task 4: Sentiment analysis in twitter, in ‘Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)’, pp. 502–518.

KIM, Y., 2014. ‘Convolutional neural networks for sentence classification’, arXiv preprint arXiv:1408.5882 .

SOCHER, R., PERELYGIN, A., WU, J., CHUANG, J., MANNING, C. D., NG, A. Y. and POTTS, C., 2013. Recursive deep models for semantic compositionality over a sentiment treebank, *in* ‘Proceedings of the 2013 conference on empirical methods in natural language processing’, pp. 1631–1642.

WANG, X., LIU, Y., SUN, C.-J., WANG, B. and WANG, X., 2015. Predicting polarities of tweets by composing word embeddings with long short-term memory, *in* Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)’, pp. 1343–1353.

YIN, D., MENG, T. & CHANG, K.-W., 2020. ‘Sentibert: A transferable transformer-based architecture for compositional sentiment semantics’, *arXiv preprint arXiv:2005.04114* .

CAMBRIA, E., PORIA, S., BAJPAI, R. and SCHULLER, B., 2016. Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives, *in* ‘Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers’, pp. 2666–2677.

GIEVSKA, S., KOROVESHOVSKI, K. and CHAVDAROVA, T., 2014. A hybrid approach for emotion detection in support of affective interaction, *in* ‘Data Mining Workshop (ICDMW), 2014 IEEE International Conference on’, IEEE, pp. 352–359.

RECUPERO, D. R., PRESUTTI, V., CONSOLI, S., GANGEMI, A. and NUZZOLESE, A. G., 2015. ‘Sentilo: frame-based sentiment analysis’, *Cognitive Computation* **7**(2), 211–225.

YU, L.-C., WANG, J., LAI, K. R. and ZHANG, X., 2017. Refining word embeddings for sentiment analysis, *in* ‘Proceedings of the 2017 conference on empirical methods in natural language processing’, pp. 534–539.

GILBERT, C. and HUTTO, E., 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text, in ‘Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) [http://comp.](http://comp/) social. gatech. edu/papers/icwsm14. vader. hutto. pdf’, Vol. 81, p. 82.

KIRITCHENKO, S., Zhu, X., Cherry, C., and Mohammad, S. (2014a). Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 437–442.

LIU, B. 2011. Opinion mining and sentiment analysis. In Web Data Mining, pages 459–526. Springer.