**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 |
| ISIC | The International Skin Imaging Collaboration |
| MEL | Melanoma |
| NV | Melanocytic nevus |
| VASC | Vascular lesion |

# 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.