T.C.
HARRAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

SOSYAL MEDYA VERİLERİ ÜZERİNDE DERİN ÖĞRENME TABANLI HİBRİT MODEL KULLANILARAK SALGIN DÖNEMİ DUYGU ANALİZİ

Ayhan AKKAYA

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

ŞANLIURFA
2023

İÇİNDEKİLER

Sayfa No

ÖZET i
ABSTRACT ii
TEŞEKKÜR iii
ŞEKİLLER DİZİNİ iv
ÇİZELGELER DİZİNİ v
SİMGELER ve KISALTMALAR DİZİNİ vi
1. GİRİŞ 1
1.1. Çalışmanın İçeriği 4
2. ÖNCEKİ ÇALIŞMALAR 5
3. MATERYAL ve YÖNTEM 10
3.1. Materyal 10
3.1.1. Veri toplama 10
3.1.1.1. Reddit 10
3.1.2. Verilerin çekilmesi 11
3.1.3. Verilerin temizlenmesi 13
3.2. Yöntem 17
3.2.1. Duygu polarite sınıflandırması 17
3.2.2. Vader 18
3.2.3. sparknlp 20
3.2.4. Bert 21
3.2.5. Değerlendirme metrikleri 24
4. ARAŞTIRMA BULGULARI ve TARTIŞMA 26
4.1. 3 Sınıflı Duygu Sınıflandırma Ve Eğitim Sonuçları 26
4.2. 4 Sınıflı Duygu Sınıflandırma Ve Eğitim Sonuçları 29
4.3. Aşı Ve Aşı Markaları İle İlgili Duygu Analizi Çalışmaları 31
4.3.1. 3 sınıflı duygu analiz çalışmaları 32
4.3.2. 4 sınıflı duygu analiz çalışmaları 35
5. SONUÇLAR ve ÖNERİLER 39
KAYNAKLAR 40
ÖZGEÇMİŞ 42

ÖZET

Yüksek Lisans Tezi

SOSYAL MEDYA VERİLERİ ÜZERİNDE DERİN ÖĞRENME TABANLI HİBRİT MODEL KULLANARAK SALGIN DÖNEMİ DUYGU ANALİZİ

Ayhan AKKAYA

Harran Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Nagehan İLHAN
YIL: 2023, Sayfa: 42

Sosyal medyanın gün geçtikçe insan hayatında daha fazla önem kazandığı hatta birçok konuda adeta karar organı haline geldiği görülmektedir. Özellikle 2019 yılı sonlarında hayatımızı etkisi altına alan koronavirüs salgını insanları toplumsal anlamda birbirinden daha da uzaklaştırmıştır. Koronavirüs salgınının pandemi haline gelmesinden sonra oluşan her gelişme insanların en önemli gündem maddelerinden biri olmuştur. Salgına son vereceği düşünülen aşılar hakkında insanların düşünceleri ve duyguları mutlaka bilinmesi gereken bir konu haline gelmiştir. 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. Bu çalışmada bu düşünceden yola çıkarak öncelikle insanların en çok yoğunlaştığı konu başlıklarından 2020 ve 2021 yılındaki veriler çekilmiştir. Verilerin doğru duygu etiketlerine sahip olması için veriler üzerinde birçok veri ön işleme işlemleri gerçekleştirilmiştir. Çok sınıflı duygu etiketlemesi yapmak için birden fazla modeli üst üste kullanarak hibrit bir model oluşturulmaya çalışılmıştır. Sosyal medya gönderileri hem pozitif, negatif, nötr ve aynı zamanda korku, mutsuzluk, eğlenceli, mutlu şeklinde farklı sınıflandırıcıları kullanarak etiketlenmiştir. Yüksek başarımlı eğitim işlemleri sonuçlarına göre insanların aşı ve aşı markaları ile ilgili düşüncelerine odaklanılmıştır. Çalışmanın başarım oranlarını yükseltmek için derin öğrenme tabanlı bazı sınıflandırıcılar kullanılmıştır. Yapılan bu çalışmanın sosyal medya verilerinin sınıflandırılması ile ilgili farklı alanlarda başarılı olabileceği varsayılmaktadır.

ANAHTAR KELİMELER: Duygu analizi, sosyal medya, aşı, koronavirüs, doğal dil işleme

.

ABSTRACT

MSc Thesis

EPIDEMIC PERIOD EMOTION ANALYSIS USING A DEEP LEARNING BASED HYBRID MODEL ON SOCIAL MEDIA DATA

Ayhan AKKAYA

Harran University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering

Supervisor: Assist. Prof. Dr. Nagehan İLHAN
Year: 2023, Page No: 42

It is seen that social media is gaining more importance in human life day by day and has even become a decision-making body on many issues. Especially the coronavirus epidemic, which affected our lives at the end of 2019, has pushed people further away from each other socially. Every development that has occurred after the coronavirus epidemic has become a pandemic has been one of the most important agenda items of people. People's thoughts and feelings about vaccines, which are thought to end the epidemic, have become a subject that must be known. Perhaps the biggest proof of this is that the topics opened on this subject at the beginning of the pandemic are still up to date. It is as important to make the right meaning of these emotions as it is to obtain them. In this study, based on this idea, first of all, the data in 2020 and 2021 were taken from the topics that people focused on the most. In order for the data to have the correct emotion labels, many data preprocessing operations were performed on the data. We tried to create a hybrid model by using multiple models on top of each other to perform multi-class sentiment labeling. Social media posts were labeled using different classifiers such as positive, negative, neutral and also fear, sadness, fun, happy. According to the results of high-performance training operations, we focused on people's opinions about vaccines and vaccine brands. Some deep learning-based classifiers were used to increase the success rates of the study. It is assumed that this study can be successful in different areas related to the classification of social media data.

KEY WORDS: Sentiment analysis, social media, vaccine, coronavirus, natural language processing

TEŞEKKÜR

Akademik Kariyerimi geliştirmek adına çıktığım bu zor yolculukta alanında engin bilgi ve tecrübelerini benimle paylaşan ve her konuda bana desteğini bir an olsun esirgemeyen değerli danışman hocam Dr. Öğr. Üyesi Nagehan İLHAN'a, deneysel çalışmalarımızı daha sağlıklı yapabilmek adına bize yüksek özellikli bilgisayarına erişim sağladığı için Doç. Dr. Mehmet Emin TENEKECİ'ye ve bu zorlu süreçte dualarını bir an olsun bırakmayan anneme, babama ve her zaman yanımda olan beni destekleyen, bana inanan sevgili eşim Meral ve kızım Emine'ye teşekkür ederim.

ŞEKİLLER DİZİNİ

Sayfa No

Şekil 1.1.Duygu Sınıflandırma Yöntemleri. 3
Şekil 3.1.Reddit Pushshift Veri Toplama Platformu. 11
Şekil 3.2.2020 Yılı Aylara Göre Çekilen Veri Sayıları. 12
Şekil 3.3.2021 Yılı Aylara Göre Çekilen Veri Sayıları. 12
Şekil 3.4.Veri Temizleme Aşamaları 15
Şekil 3.5.Verisetinde En Çok Bulunan 50 Kelime 17
Şekil 3.6.sparkNLP Çalışma Yapısı 20
Şekil 3.7.BERT Girdi Katmanı. 22
Şekil 3.8.Duygu Sınıflandırmada kullanılan BERT Mimarisi 23
Şekil 4.1.Etiketleme ve Eğitim İşlemleri Sonrası Oluşan 3 Sınıflı Duygu Oranları 28
Şekil 4.2. Etiketleme ve Eğitim İşlemleri Sonrası Oluşan 4 Sınıflı Duygu Oranları 31
Şekil 4.3. Aşı ve Aşı Markalarına İlişkin 3 Sınıflı Duygu Değişimleri 32
Şekil 4.4. 2020 Yılı Aylara Göre Duygu Değişimleri 33
Şekil 4.5. 2021 Yılı Aylara Göre Duygu Değişimleri 34
Şekil 4.6. Aşı ve Aşı Markalarına İlişkin 4 Sınıflı Duygu Değişimleri 35
Şekil 4.7. 2020 Yılı Aylara Göre 4 Sınıflı Duygu Değişimleri 37
Şekil 4.8. 2021 Yılı Aylara Göre 4 Sınıflı Duygu Değişimleri 38

ÇİZELGELER DİZİNİ

Sayfa No

Çizelge 3.1 PushShift'den çekilen verilerin alanlarına ait bilgiler 13
Çizelge 3.2.Temizlenmiş verilere ait bazı örnekler 16
Çizelge 4.1.Vader ile 3 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri 27
Çizelge 4.2.Glove ile 3 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri 27
Çizelge 4.3.Bert ile 3 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri 28
Çizelge 4.4.sparkNLP ile 4 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri 29
Çizelge 4.5.Glove ile 4 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri 30
Çizelge 4.6. BERT ile 4 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri 30

SİMGELER ve KISALTMALAR DİZİNİ

API	Application Programming Interface (Uygulama Programlama Arayüzü)
BERT	Bidirectional Encoder Representations from Transformers (Transformatörlerden Çift Yönlü Kodlayıcı Temsilleri)
CNN	Convolutional Neural Network (Evrişimli Sinir Ağı)
GLOVE	Global Vectors for Word Representation (Kelime Temsili için Global Vektörler)
LIWC	Linguistic Inquiry Word Count (Dilbilimsel Sorgulama Kelime Sayısı)
NLP	Natural Language Processing (Doğal Dil İşleme)
NP	Noun Phrase (İsim Cümlesi)
POST	Parts Of Speech Tagging (Konuşma Etiketlemenin Bölümleri)
RNN	Recurrent Neural Network(Tekrarlayan Sinir Ağı)
SVM	Support Vector Machine (Destek Vektör Makinesi)
VADER	Valence Aware Dictionary for Sentiment Reasoning (Duygu Akıl Yürütme için Değer Duyarlı Sözlük)
VP	Verb Phrase (Fiil Cümlesi)

1. 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; İstatistik, Veritabanı Yönetim sistemleri ve Makine Öğrenimi konularının kesiştiği noktadaki yöntemleri içeren büyük veri kümelerindeki kalıpları ortaya çıkarma 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 getirilip 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 gelmektedir. 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 halini almaktadır.

Reddit, bir Amerikan kaynaklı sosyal medya haber toplama, tartışma ve web sitesi içeriği derecelendirme sitesidir. Reddit' e kayıtlı üyeler; sayfa içi bağlantılar, metin gönderileri ve resimler vb. içerikleri paylaşır ve bu içerikler sonradan 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 sağlamaktadır. Hatta bununla ilgili uygulamalar ve yazılım dillerine ait kütüphaneler bile yazılmıştır.

Koronavirüs salgını (COVID-19), diğerlerinden farklı olarak sonradan ortaya çıkan bir koronavirüs türünün sebep olduğu bulaşıcılık içeren bir hastalıktır (Singhal, 2020). 2019 yılı sonlarında Çin'in Wuhan şehrinde bir balık pazarında ilk olarak görüldüğü düşünülen bu hastalık başlangıçta o bölgede etkili olsa da ilerleyen süreçte global anlamda tüm dünyayı etkileyen 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 yapılmıştır. 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 iç görü 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 değerlilik (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 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 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(Deng, 2014). Aşağıda Şekil 1.1.'de Duygu sınıflandırma yöntemlerine ilişkin şema verilmiştir.



Şekil 1.1. Duygu Sınıflandırma Yöntemleri

1.1. Çalışmanın İçeriği

Bu çalışmada reddit sosyal medya platformu kullanılarak 2020 yılı başlarından itibaren 2021 yılı sonuna kadar sosyal medyanın aşı ve aşı markaları hakkındaki verileri sosyal medya platformu olan Reddit'ten çekilmiştir. Verileri aylık olarak çekmek için Reddit'ten veri çekmeye olanak sağlayan Python programlama dilinde yazılmış PSAW Uygulama Programlama Arayüzü (Application Programming Interface -API) kullanılmıştır. Reddit platformunda çekilen yorumlardan İngilizce yorumlar üzerinde duygular etiketlemek için sözlük tabanlı bazı sınıflandırıcılar kullanılmıştır. 3 sınıflı (Positive – Negative – Neutral) duygu etiketleme işlemi için Duygu Akıl Yürütme için Değer Duyarlı Sözlük (Valence Aware Dictionary for Sentiment Reasoning-VADER) sınıflandırıcı kullanılmıştır. 4 Sınıflı (Fear – Sadness – Joy – Surprise) duygu etiketleme işlemi için ise sparkNLP sınıflandırıcısı kullanılmıştır. Daha sonra verilerin doğruluğunu arttırmak için hem 3 sınıflı hem de 4 sınıflı veriler üzerinde kelime temsillerini ifade eden kelime gömme yapısı olan Kelime Temsili için Global Vektörler (Global Vectors for Word Representation-Glove) ön eğitimli modeli verilere uygulanmıştır. Glove ile eğitim işlemleri sonucunda 3 sınıflı veriler için %75 ve 4 sınıflı veriler için %72 gibi bir doğruluk oranı elde edilmiştir. Daha yüksek bir doğruluk oranına ulaşmak için Glove ile eğitilen veriler üzerinde son zamanlarda özellikle Doğal Dil İşleme (Natural Language Processing-NLP) alanında kullanılan Derin Öğrenme tabanlı bir kelime gömme yapısı olan Transformatörlerden Çift Yönlü Kodlayıcı Temsilleri (Bidirectional Encoder Representations from Transformers-BERT) ile eğitim işlemi gerçekleştirilmiştir. Bert ile eğitim işlemleri sonucunda 3 sınıflı veriler için %82 ve 4 sınıflı veriler için % 78 gibi bir başarı oranı yakalanmıştır. Ham olarak çekilen veriler üzerinde yapılan bir dizi işlem sonucunda veriler daha anlamlı bir hale gelmiştir. Çalışmada odaklanılan asıl amaç farklı markalara ve aşıya ait insanların duygularını analiz edip insanlara somut bir çalışma sunmaktır.

2. ÖNCEKİ ÇALIŞMALAR

Duygu analizi ile ilgili farklı alanlarda yapılmış birçok çalışma bulunmaktadır.

NLP, Bilgisayar sitemlerinin konuşma dillerinde yazılmış ifadeleri veya kelimeleri anlamasını sağlamaya adanmış bir Yapay Zeka ve Dilbilim alanıdır (Khurana, 2017). NLP'nin birçok çalışmada kullanılan görevlerinden bazıları Adlandırılmış Varlık Tanıma, Konuşma Parçaları Etiketleme, Parçalama, Stopwords kaldırma, Lemmatization vb. metindeki öğeler özel adlarla ilgilidir. Konuşma Etiketlemenin Bölümleri (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 (Noun Phrase-NP) ve Fiil Cümlesi (Verb Phrase-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, Lemmatizasyon görevi kullanılarak yapılabilir. Lemmatizasyon kavramı normalde kelimelerde bulunan çekim eklerini kaldırmayı amaçlar. Bu işlemi yaparken de lemma olarak adlandırılan bir kelimeye ait sözlük formunu döndürür (Manning, 2009). Bu tür tekniklerin kullanımı, algoritmaları eskisinden daha kolay hale getirmiştir.

Liu'ya göre duygu analizi kavramı, insanların ürünler, bireyler, organizasyonlar ve olayların nitelikleri le ilgili fikirleri, tutumları ve duygu durumlarını analiz eden bir kavramdır. Bu 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 çıkarır) (Bing, 2012).

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, Maksimum Entropi Sınıflandırması, Naive Bayes ve Destek Vektör Makineleri (Support Vector Machine-SVM) (Bo, 2002) olmak üzere üç makine öğrenme modeli üzerinde eğitilmiştir. Bu araştırmaya dayanarak, derin öğrenme sistemlerinin geliştirilmesine yönelik çalışmalar genişletilmiştir.

Dong ve Wei, 2014'te Twitter verilerini örnek olarak kullanarak duygu sınıflandırması için uyarlanabilir özyinelemeli Sinir Ağı (AdaRNN) önermiştir (Dong, 2014). Derin öğrenme tekniklerindeki artışla birlikte, Twitter verileri üzerinde duygu analizi yapmak için Coooolll adlı bir derin öğrenme sistemi geliştirilmiştir. Duyguları olumlu ve olumsuz kategorilere ayırmak için %87,61 doğruluk hesaplamışlardır (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 çalışmış ve hem eski hem de yeni etiketlenmemiş verileri kullanan bir yöntem önermişlerdir. Duygu sınıflandırması için Bulanık derin inanç ağları adı verilen iki aşamalı yarı denetimli bir öğrenme yöntemi önermişlerdir. Model, film incelemeleri ve DVD incelemeleri veri setinde test edilmiş ve %79.5'lik bir doğruluk elde edilmiştir (Zhou, 2014).

Sözlük tabanlı yaklaşımda, 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" denilmektedir (Palanisamy, 2013).

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 SVM gibi geleneksel makine öğrenme tekniklerini hem de derin öğrenme modellerini kapsamaktadır. 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 Evrişimli Sinir Ağları (Convolutional Neural Network-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 (Long-Short Term Memory-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 odaklanmışlardır. 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 kullanmıştır (Recupero, 2015).

Yu ve arkadaşları, önceden eğitilmiş kelime yerleştirmelerine GloVe uygulanabilecek bir vektör iyileştirme modeli önermiştir. Burada mevcut kelime

yerleştirmeleri, anlamsal ve duygusal olarak benzer kelimeler birbirine daha yakın olacak ve bunun tersi olacak şekilde ayarlanmıştır. Bu, gerçek değerli duygu puanlarını içeren bir duygu sözlüğü kullanılarak yapılmıştır (Yu, 2017).

3. MATERYAL ve YÖNTEM

3.1. Materyal

3.1.1. 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.

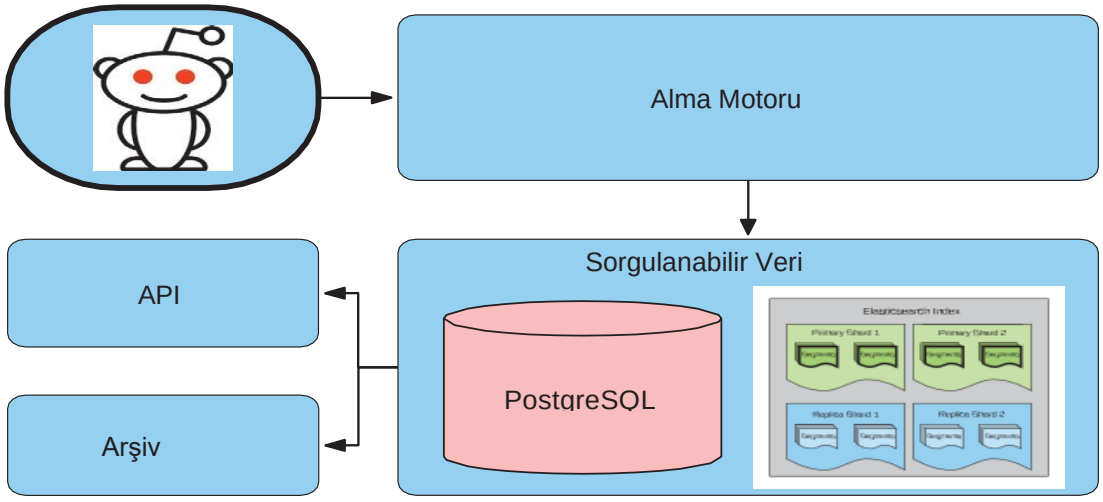
3.1.1.1. Reddit

Web verileri, özellikle API'lerden gelen veriler, çevrimiçi sosyal platformların kullanıcı tarafından oluşturulan etkinlik ve içerik veritabanlarını kullanan araştırmacılar için muazzam bir nimet olmuştur. 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 3.1.'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.

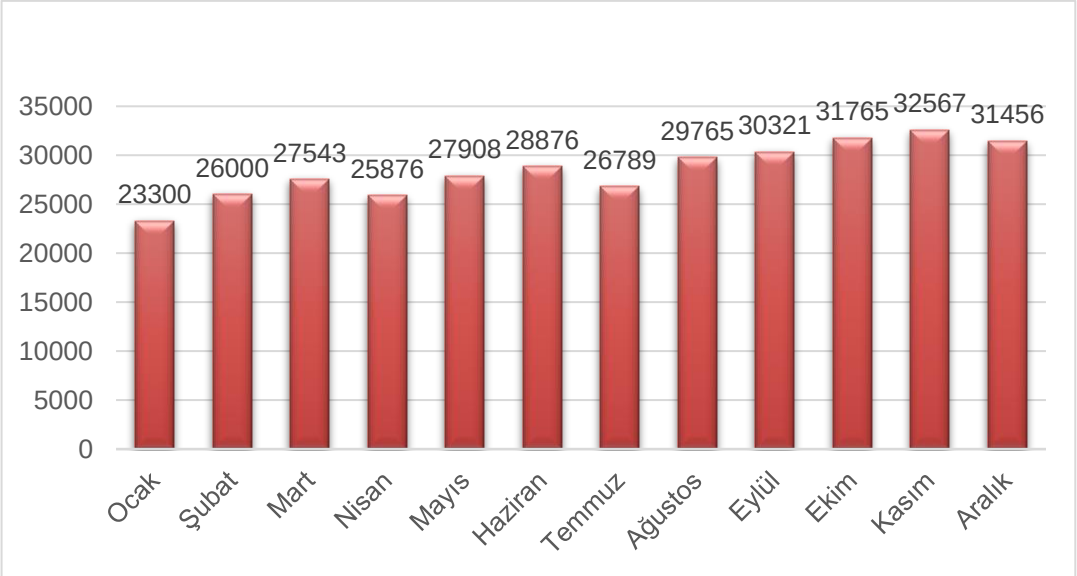
Alınan verilerin indekslenmesini ve toplanmasını gerçekleştiren bir Elastic Search belge deposu kümesi



Şekil 3.1. Reddit Pushshift Veri Toplama Platformu.

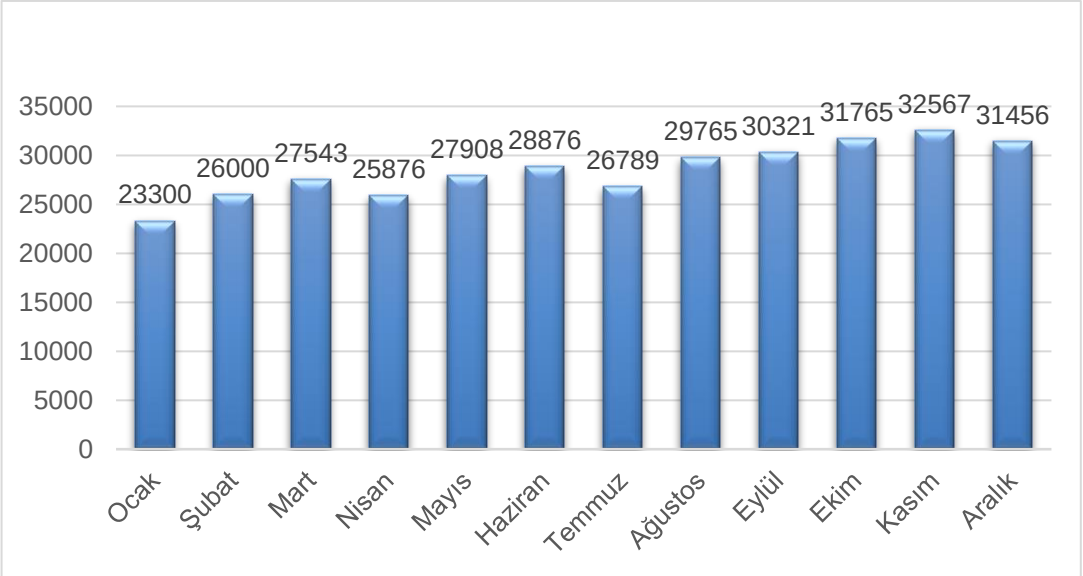
3.1.2. Verilerin çekilmesi

Verileri çekmek için DataExtract adında aylara göre belirlenen alt konu başlıklarında verilerin çekilmesine olanak sağlayan Python programlama dilinde bir sınıf tanımlaması yapılmıştır. Konu başlıkları, ay sayısı gibi verilerin dışarıdan girilmesine olanak sağlayan bu sınıf sayesinde farklı alanlarda da verilerin çok daha kolay bir şekilde çekilmesine olanak sağlanmıştır. Ayrıca veriler çekilirken kullanıcılara ait yorumlar ayrı bir json dosyasında yedeklenmiştir. Bu şekilde kullanıcıların da tüm yorumlarının genel bir duygu sınıflandırmalarına ilişkin bilgi edinilmiştir. Kullanıcı yorumlarından ortalama duygular çıkarılıp kullanıcıların aylara göre duygu değişimlerini görme fırsatı da elde edilmiştir. Şekil 3.2.'de 2020 yılına ait, Şekil 3.3.'de ise 2021 yılına ait aylara göre çekilen veri miktarları gösterilmiştir. 2020 yılı başlarında pandeminin insan hayatına yeni girmesinden dolayı çok fazla veri bulunamadığı için sonraki ayalara göre daha az olmuştur.



Şekil 3.2. 2020 Yılı Aylara Göre Çekilen Veri Sayıları

2021 yılında ise pandemi adeta hayatın bir parçası haline gelmiştir. Aşı ve virüs ile ilgili farklı alt başlıklardan çok daha fazla veri çekilebilmiştir. Aylara göre benzer oranlarda veri çekilmesi duygu değişimlerinin daha gerçekçi izlenilebilmesi anlamında önem arz etmiştir.



Şekil 3.3. 2021 Yılı Aylara Göre Çekilen Veri Sayıları

Çizelge 3.1.'de PushShift API ile çekilen verilere ait bazı alanlar ve açıklamaları verilmiştir. Özellikle selftext alanındaki yorumlar üzerinde işlemler gerçekleştirilmiştir.

Çizelge 3.1. PushShift'den çekilen verilerin alanlarına ait bilgiler

Alan	Açıklama
id	The gönderiyi yapan, örneğin., "3lhg2"
url	Gönderinin gönderildiği internet adresi bilgisi. Örneğin, "https://www.reddit.com/r/Coronavirus/
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

3.1.3. 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 söz dizimsel ve anlamsal belirsizlikler, yapılan yorumlar üzerinde gerçekleştirmek istenilen 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 (Stopwords) : 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 Dönüştürme (Lowercase) : Bu terminoloji, bir metindeki tüm kelimelerin küçük harfe dönüştürülmesini ifade eder. Esas olarak girdinin tutarlılığı ve uyumu nedenleriyle benimsenmiştir.

Köklerine Ayırma (Tokenization) : Bu ifade, bir metnin, köken 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 belirtilen DataExtract isimli sınıf içerisinde yukarıda tanımlanan veri temizleme işlemleri dataProcessing isimli fonksiyon ile tanımlanmıştır. Böylece veriler çekilirken aynı zamanda temizleme işlemi de gerçekleştirilmiştir.



Şekil 3.4. Veri Temizleme Aşamaları

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

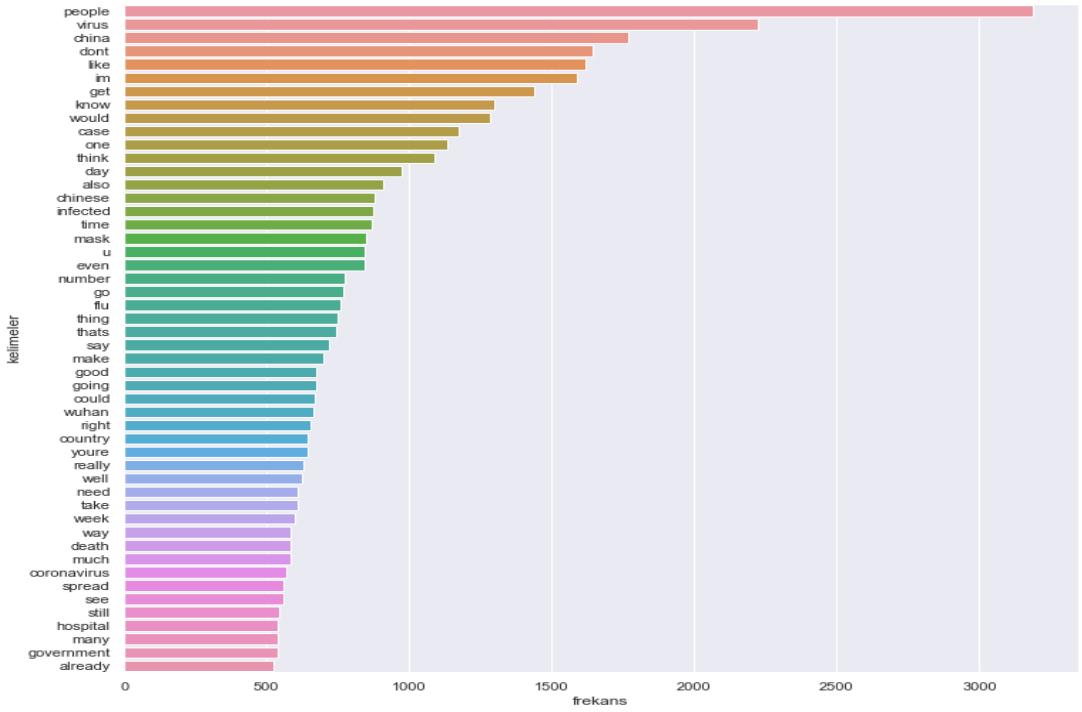
Veriler çekilip temizlendikten sonra üzerinde hızlıca sınıflandırma ve eğitim işlemleri yapabilmek için clean_data.db isminde bir veritabanında veriler tablo olarak tutulmuştur. Bütün bu işlemler sonrasında ise artık veriler sınıflandırma işlemleri için hazır hale gelmiştir.

Çizelge 3.2.'de temizlenmiş verilere ilişkin farklı tarihlerde farklı subreddit içerisinden bazı örnekler verilmiştir. Çalışma İngilizce veriler üzerinde etkili olduğu için ingilizce örnekler seçilmiştir. Çizelge 3.2.'de clean_text sütununda verilerin eğitim işlemlerine hazır hale getirilmiş birkaç örneği görülmektedir. Üstelik farklı zamanlara ait bazı yorumlar eklenmiştir.

Çizelge 3.2. Temizlenmiş verilere ait bazı örnekler

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

Veriler çekilip ön işleme sokulduktan sonra en çok kullanılan kelimeler Şekil 3.5.'te gösterilmiştir. İngilizce içerikler üzerinden hareket edildiği için en çok çıkan kelimelerde aşı ile ilgili birçok kelimenin İngilizce olarak görülmektedir. Şekil 3.5.'e göre vaccine (aşı), people(insanlar), virüs(virüs), china(çin) gibi kelimelerin de bu aralıkta olduğu dikkat çekmektedir. Bu durum da aslında çekilen verilerin ilgili konuyla paralel yönde seyrettiğini göstermektedir.



Şekil 3.5. Verisetinde En Çok Bulunan 50 Kelime

3.2. Yöntem

Bu bölümde verilerin etiketlenmesi, eğitilmesi gibi bir takım işlemler hakkında bazı bilgiler verilmiştir. Bu işlemler sonrası veriler artık üzerinde analiz işlemleri yapılabilecek hale gelmiş olacaktır. Aşağıda uygulanan yöntemler ile ilgili bilgiler yer almaktadır.

3.2.1. Duygu polarite sınıflandırması

Duygu polarite sınıflandırması, duygu analizi alanında en çok çalışılan problemlerden biridir. Subjektif bir metin verildiğinde amaç, metnin genel tonunun olumlu, olumsuz veya nötr olup olmadığını belirlemektir. Sözcüklerin ve ifadelerin duygu kutupluluğu ilişkilendirmelerinden, cümlelerin, SMS mesajlarının, sohbet mesajlarının ve tweet'lerin duyarlılığına, ürün incelemelerinde, blog gönderilerinde ve bütünlükteki duygu analizine kadar tüm düzeylerde gerçekleştirilir.

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. 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).

3.2.2. Vader

Vader, 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 sergilemektedir. Vader, Dilbilimsel Sorgulama Kelime Sayısı (Linguistic Inquiry Word Count-LIWC) 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 anlam bakımından duygu yönelimlerine göre olumlu ya da olumsuz olarak etiketlenen sözcüksel özelliklerin bir listesi olan duygu sözlüğünün bir kombinasyonunu kullanır. Vader, sosyal medya metinleri, film ve dizi incelemeleri ve ürün incelemelerinde başarılı olmuştur. Bunun nedeni, Vader'in yalnızca olumluluk ve olumsuzluk puanını değil, bir duygunun ne kadar olumlu veya ne kadar olumsuz olduğunu da söylemesidir. Vader tasarımcıları, bu derecelendirme puanlarının birçoğunda Amazon'un Mechanical Türk yapısını kullanmışlardır.

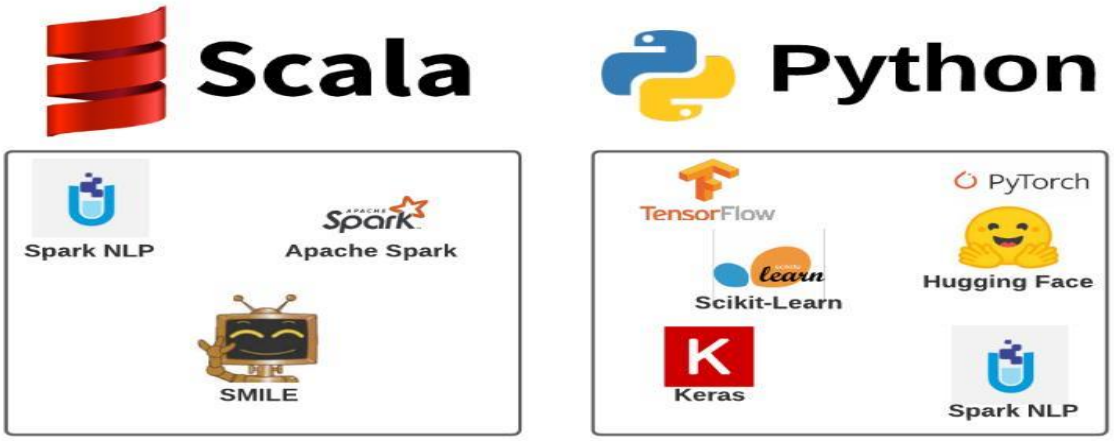
Vader, metindeki kelimelerden herhangi birinin Vader sözlüğünde olup olmadığını görmek için bir metin parçasını analiz eder. polarity_scores() fonksiyonunu kullanarak polarite indekslerini bulabilir. Bu, belirli bir cümle için negatif, nötr, pozitif ve bileşiğin metrik değerlerini döndürür. Bileşik puan, -1 ile +1 arasında normale indirgenmiş sözlük derecelendirme puanlarının toplam değerini hesaplayan bir ölçüm sistemidir; 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.

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 (Hutto, 2014).

3.2.3. Sparknlp



Şekil 3.6. sparkNLP Çalışma Yapısı

sparkNLP, Apache Spark ve SparkML üzerine inşa edilmiş açık kaynaklı bir metin işleme kod kütüphanesidir. Doğal Dil İşleme ardışık düzenleriyle kullanılabilecek bir API sağlar. Kütüphane, özel modeller oluşturmanın yanı sıra eğitimli modeller ve işlem hatları kullanmak için destek sağlar.

sparkNLP çerçevesi, farklı dil temsil modelleri sağlar. Desteklenen modeller, çeşitli metin sınıflandırma görevlerini gerçekleştirmek için yeniden kullanılabilir. sparkNLP, iki model kullanarak duyarlılık analizinin uygulanmasını mümkün kılmaktadır. Bunlar: SentimentDLApproach ve ClassifierDLApproach modelleridir.

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.

3.2.4. Bert

Bert, Google AI ekibi tarafından 2018'de derin öğrenme tekniklerine dayalı olarak önerilen önceden eğitilmiş bir dil temsil modelidir (Devlin, 2019). Diğer dil temsil modellerinden farklı olarak, katmanlarında bağlamı hem sol hem de sağ taraftan 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, 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 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 sağlamaktadır.

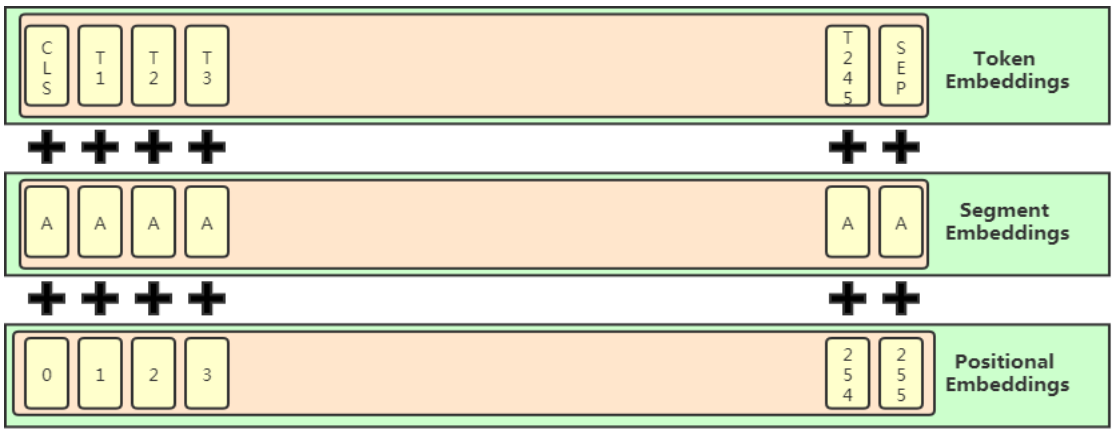
Maskeli Dil Modeli Amaç: Model, A dizisi ve B dizisinde rastgele maskelenen belirteçleri tahmin etmeyi öğrenir.

Sonraki Cümle Tahmini: Bert'in 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, transformatör tekrarlayan veya evrişimli sinir ağları yerine temel bileşenler olarak kullanılır. Transformatör yalnızca kendi kendine dikkat mekanizmasına dayanmaktadır. Tekrarlayan Sinir Ağı (Recurrent Neural Network-RNN) veya 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'e bir ek çıktı katmanıyla ince ayar yaparak mükemmel bir performans modeli oluşturmak kolaydır (Devlin, 2019).



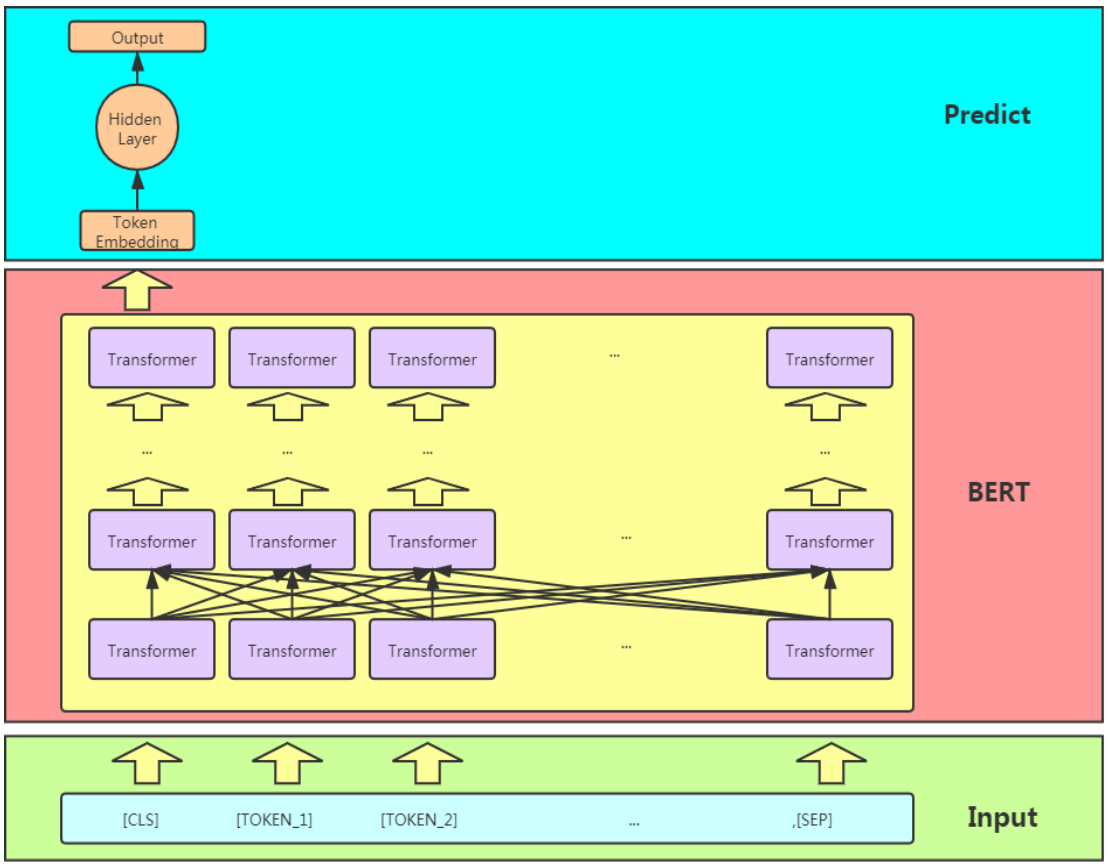
Şekil 3.7. BERT Girdi Katmanı (Devlin, 2019)

Google, Github'da önceden eğitilmiş BERT'i farklı boyutlarda güncellemiştir. Çalışmalar 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 3.8.'de aşağıdaki üç parça ile sunulmaktadır.

Giriş: İlk kısım giriş katmanıdır. Sorumluluğu, ilk gömme kelimesini kabul etmek ve onu Bert'e teslim etmektir.

BERT: $B\text{ert}(\cdot)$ önceden eğitilmiş Bert modeli olsun, x giriş kısmından ilk gömme olsun. Önce gizli gösterimi $Z = B\text{ert}(x) \in Rrh*|x|$ olarak elde ederiz, burada $|x|$ giriş dizisinin uzunluğu ve rh gizli boyutun uzunluğudur. Bu bölümün çıktısı, her bir girdi belirtecinin son sözcük gömmesidir.

Tahmin: Gizli gösterim yoğun bir katmana ve ardından $softmax$ fonksiyonlarına iletilir: $g = softmax(w \cdot Z + b) = softmax(B\text{ert}(x))$, burada $W \in Rrh\ and\ b \in R$. $softmax$, 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 $\in$ Rk ve k, veri kümesinde negatif, nötr ve pozitif dahil olarak üç duygu olduğu için 3'e eşittir.



Şekil 3.8. Duygu Sınıflandırmada kullanılan Bert Mimarisi (Devlin, 2019)

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.

3.2.5. Değerlendirme metrikleri

Derin öğrenme modellerinde ilk olarak, bir model yapısı oluşturulur ve modelin ölçümleri ile geri bildirim alınır. Ardından, gerekli iyileştirmeler yapılır ve istenilen başarı oranına ulaşana kadar çalışmalar devam eder. Bu süreç, farklı modellerle tekrar edilir ve her bir deneyin sonucu, ölçme metrikleri kullanılarak değerlendirilir.

Duygu analizi için eğitilen modelin test edilmesinde, çeşitli standart değerlendirme ölçütleri kullanılır. Bu ölçütler arasında aşağıdaki denklemlerde verilen doğruluk(accuracy), kesinlik(precision), geri çağırma(recall) ve F ölçümü (F1-Score) denilen değerlendirme metrikleri bulunur. Bu metrikler, bir modelin performansını açıklamanın yanı sıra sonuçları ayırt etme yeteneklerini de ölçerler.

Özellikle doğruluk, bir modelin ne kadar doğru sonuçlar verdiğini ölçerken, kesinlik bir modelin ne kadar doğru pozitif sonuçlar verdiğini ölçer. Geri çağırım veya hatırlama ise, bir modelin ne kadar gerçek pozitif sonuçları yakalayabildiğini ölçer. F ölçümü ise, hem kesinlik hem de geri çağırım ölçümlerini birleştirerek bir modelin performansını ölçer.

Bu değerlendirme ölçütleri, bir modelin performansını ayrıntılı bir şekilde ölçerken, modelin geliştirilmesi ve iyileştirilmesi için de kullanılabilirler. Örneğin, doğruluk oranını arttırmak için modeldeki değişkenler ayarlanabilir, kesinlik oranını arttırmak için veri setleri daha dikkatli seçilebilir veya geri çağırım oranını arttırmak için farklı algoritmalar denenerek model iyileştirilebilir.

$$\text{Accuracy}(\%) = \left(\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}\right) \text{ x100} \quad (3.1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3.2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3.3)$$

$$\text{f1}_{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision + Recall}} \quad (3.4)$$

4. ARAŞTIRMA BULGULARI ve TARTIŞMA

Reddit sosyal medya platformundan çekilen verilerin diğer sosyal medya platformlarına nazaran verilerde gönderilerde bir karakter sınırı olmaması ve 2 yıllık verilerin çekilmesi çok büyük miktarda bir veri üzerinde eğitim işlemleri yapma zorunluluğu ortaya çıkarmıştır. Bu eğitim işlemleri için sıradan bilgisayarlar işlem yeteneği açısından yetersiz kalmıştır. Bu sebeple çalışmalarda eğitim işlemlerinin daha doğru ve yüksek hassasiyette çalışabilmesi için çalışmalar, yüksek özellikler barındıran süper bilgisayarlarda ve Google'ın geliştiricilere sağladığı colab ortamında yapılmıştır.

Bu çalışmada öncelikle veriler üzerinde veri ön işleme işlemleri gerçekleştirildikten sonra Vader yapısı kullanılarak 3 duygulu (positive – negative – neutral) hem de sparkNLP yapısı kullanılarak 4 duygulu (fear – joy – sadness – surprise) veri etiketleme işlemi gerçekleştirilmiştir. Etiketlenmiş verilerin doğruluğunu arttırmak için etiketli veriler önce Glove ile daha sonra da Bert ile ön eğitimli modeller yardımıyla eğitim işlemleri gerçekleştirilmiştir.

4.1. 3 Sınıflı Duygu Sınıflandırma Ve Eğitim Sonuçları

Etiketlenmemiş veriler üzerinde VADER yapısı kullanılarak 3 sınıflı (positive – negative –neutral) duygu sınıflandırması yapılmıştır. Çizelge 4.1.'de belirtilen sonuçlar elde edilmiştir. Çizelge 4.1.'e göre farklı duygulara ilişkin farklı değerlendirme metrikleri oluşmuştur. Doğruluk değeri 0.70 iken ortalama F1 skoru 0.71 gibi bir oran çıkmıştır.

Çizelge 4.1. Vader ile 3 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri

Sentiment (Duygular)	Precision (Hassasiyet)	Recall (Hatırlama)	F1-Score (F-Skoru)	Accuracy (Doğruluk)
Negative	0.68	0.75	0.71	0.70
Positive	0.74	0.70	0.72	
Neutral	0.69	0.64	0.67	

3 Sınıflı duygu etiketleme işlemlerinden sonra verilerin başarımlarını arttırmak için Glove ile önceden eğitilmiş modeller sayesinde veriler üzerinde bir dizi eğitim işlemleri gerçekleştirilmiştir. Eğitim işlemleri sonrasında Vader ile etiketleme işlemleri sonrasında ortaya çıkan başarım oranlarına göre daha yüksek oranlar ortaya çıkmıştır. Vader ile 0.70 bulunan Doğruluk oranları Glove ile eğitim sonrasında 0.75 gibi bir Doğruluk oranına yükselmiştir. Ayrıca Vader ile 0.71 ortalama F1 skoru oranları ortalama 0.74 gibi bir orana yükselmiştir. Ortaya çıkan tüm değerler Çizelge 4.2.'de belirtilmiştir.

Çizelge 4.2. Glove ile 3 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri

Sentiment (Duygular)	Precision (Hassasiyet)	Recall (Hatırlama)	F1-Score (F-Skoru)	Accuracy (Doğruluk)
Negative	0.72	0.79	0.75	0.75
Positive	0.76	0.74	0.76	
Neutral	0.73	0.68	0.72	

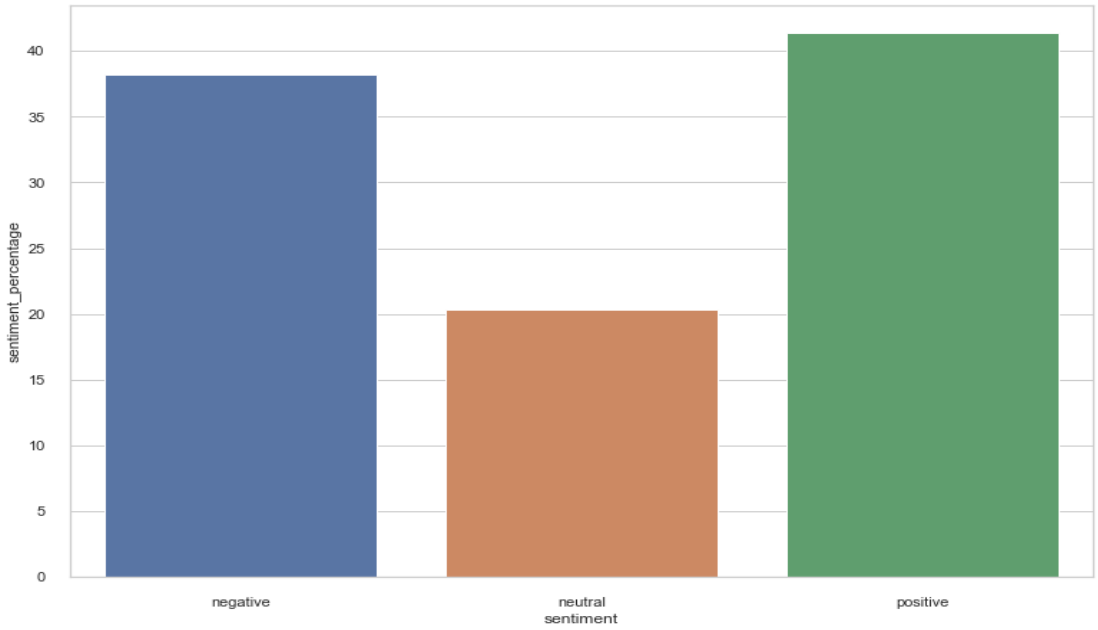
Glove yapısından sonra Bert yapısı kullanılmıştır. Bert ile önceden eğitilmiş modeller ile eğitim işlemlerine devam edilmiştir. Bert modeli ile eğitim işlemleri sonrasında Glove yapısına göre daha başarılı sonuçlar ortaya çıkmıştır.

Glove ile oluşan 0.75 Doğruluk oranları BERT ile 0.82 gibi bir Doğruluk oranına yükselmiştir. Ayrıca Glove ile oluşan ortalama 0.74 olan F1 skoru da ortalama 0.80 gibi bir orana yükselmiştir. Çalışmalar Bert modelinin Vader ve Glove a göre sınıflandırma işlemlerinde daha başarılı olduğunu göstermektedir. Çizelge 4.3.'de Bert modeli sonrası oluşan oranlar belirtilmiştir.

Çizelge 4.3. Bert ile 3 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri

Sentiment (Duygular)	Precision (Hassasiyet)	Recall (Hatırlama)	F1-Score (F-Skoru)	Accuracy (Doğruluk)
Negative	0.80	0.86	0.83	0.82
Positive	0.83	0.79	0.81	
Neutral	0.82	0.78	0.80	

3 Sınıflı duygu etiketleme ve eğitim işlemleri sonrasında ortaya çıkan yorumlara ilişkin duygu oranları Şekil 4.1.'de gösterilmiştir. Bu sonuçlara göre negatif ve pozitif duyguların birbirine yakın oranlarda olduğu gözlemlenmiştir.



Şekil 4.1. Etiketleme ve Eğitim İşlemleri Sonrası Oluşan 3 Sınıflı Duygu Oranları

4.2. 4 Sınıflı Duygu Sınıflandırma Ve Eğitim Sonuçları

Etiketlenmemiş veriler üzerinde sparkNLP yapısı kullanılarak 4 sınıflı (fear – sadness – joy – surprise) duygu sınıflandırması yapılmıştır. sparkNLP yapısı kendi içerisinde verileri sınıflandırırken classification_report metodu sayesinde değerlendirme metrikleri hesaplanmıştır. Çizelge 4.4.'de belirtilen sonuçlar elde edilmiştir. Çizelge 4.1.'e göre farklı duygulara ilişkin farklı değerlendirme metrikleri oluşmuştur. Doğruluk değeri 0.72 iken ortalama F1 skoru 0.72 gibi bir oran çıkmıştır.

Çizelge 4.4. sparkNLP ile 4 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri

Emotions (Duygular)	Precision (Hassasiyet)	Recall (Hatırlama)	F1-Score (F-Skoru)	Accuracy (Doğruluk)
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	

4 Sınıflı duygu etiketleme işlemlerinden sonra verilerin başarımlarını arttırmak için Glove ile önceden eğitilmiş modeller sayesinde veriler üzerinde bir dizi eğitim işlemleri gerçekleştirilmiştir. Eğitim işlemleri sonrasında sparkNLP ile etiketleme işlemleri sonrasında ortaya çıkan başarım oranlarına göre daha yüksek oranlar ortaya çıkmıştır. sparkNLP ile 0.72 bulunan doğruluk oranları Glove ile eğitim sonrasında 0.74 gibi bir Doğruluk oranına yükselmiştir. Ayrıca sparkNLP ile 0.72 ortalama F1 skoru oranları ortalama 0.74 gibi bir orana yükselmiştir. Ortaya çıkan tüm değerler Çizelge 4.5.'de belirtilmiştir.

Çizelge 4.5. Glove ile 4 Sınıflı Duygu Sınıflandırma Değerlendirme Metrikleri

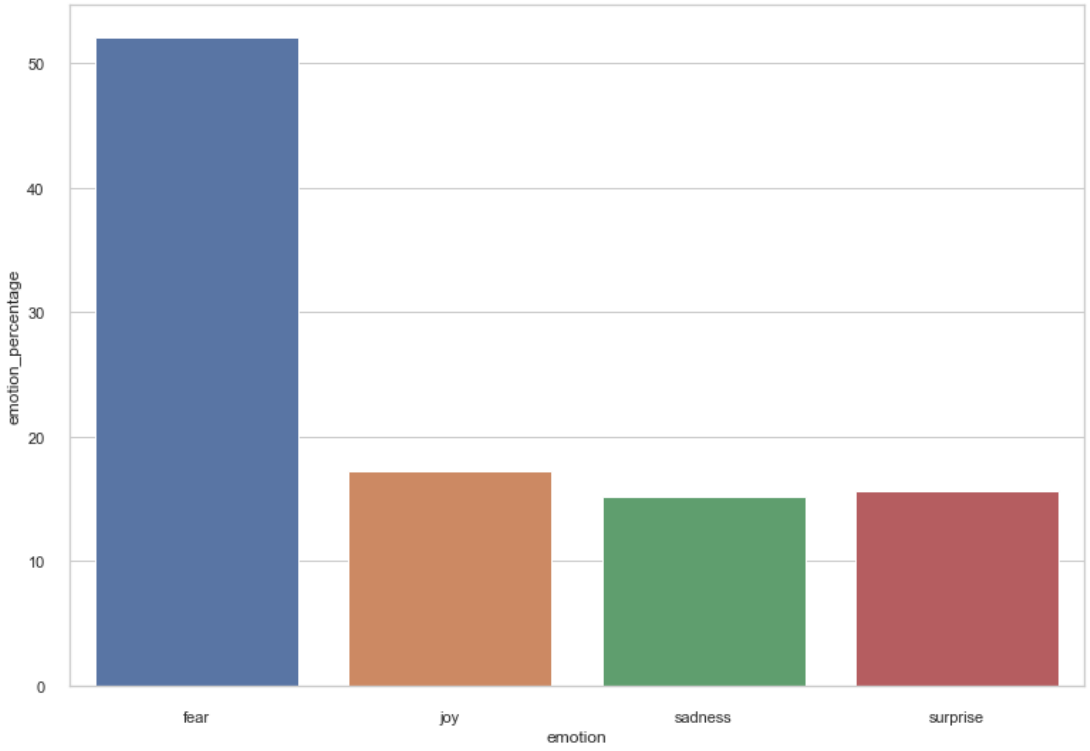
Emotions (Duygular)	Precision (Hassasiyet)	Recall (Hatırlama)	F1-Score (F-Skoru)	Accuracy (Doğruluk)
Fear (Korku)	0.73	0.77	0.74	0.74
Sadness (Mutsuzluk)	0.76	0.73	0.75	
Joy (Eğlenceli)	0.73	0.69	0.72	
Surprise (Şaşırtıcı)	0.75	0.73	0.74	

Bert ile önceden eğitilmiş modeller ile eğitim işlemlerine devam edilmiştir. Bert modeli ile eğitim işlemleri sonrasında Glove yapısına göre daha başarılı sonuçlar ortaya çıkmıştır. Glove ile oluşan 0.74 Doğruluk oranları Bert ile 0.78 gibi bir Doğruluk oranına yükselmiştir. Ayrıca Glove ile oluşan ortalama 0.74 olan F1 skoru da ortalama 0.77 gibi bir orana yükselmiştir. Çalışmalar Bert modelinin sparkNLP ve Glove'a göre sınıflandırma işlemlerinde daha başarılı olduğunu göstermektedir. Çizelge 4.6.'da Bert modeli sonrası oluşan oranlar belirtilmiştir.

Çizelge 4.6. 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	

4 Sınıflı duygu etiketleme ve eğitim işlemleri sonrasında ortaya çıkan yorumlara ilişkin duygu oranları Şekil 4.2.'de gösterilmiştir. Bu sonuçlara göre korku duygusunun insanlarda daha ağır olarak ortaya çıktığı gözlemlenmiştir.



Şekil 4.2. Etiketleme ve Eğitim İşlemleri Sonrası Oluşan 4 Sınıflı Duygu Oranları

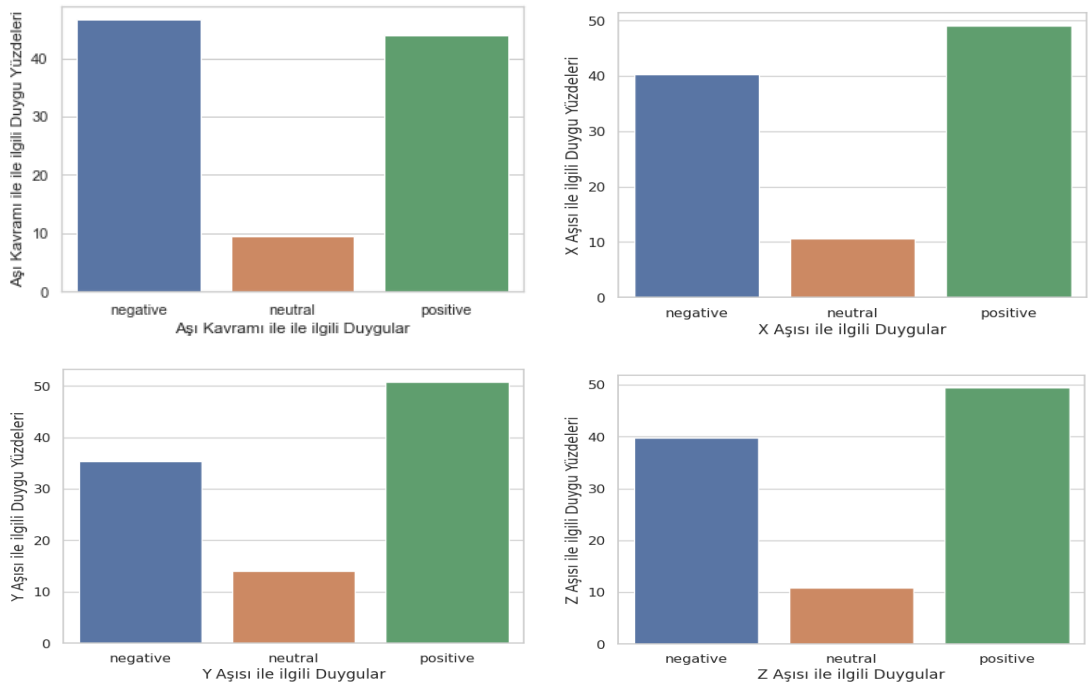
4.3. Aşı Ve Aşı Markaları İle İlgili Duygu Analiz Çalışmaları

Sosyal medya gün geçtikçe insanların hayatında daha önemli bir yer edinmektedir. Birçok konuda insanlar bazı bilgilere sosyal medya ortamından erişmektedir. Bazı konularda sosyal medyanın karar organı haline gelmiştir.

Bu bölümde eğitim işlemleri sonrasında yüksek bir başarım oranına sahip veriler üzerinde Covid-19 ile ilgili Aşı ve Aşı markalarına ilişkin bir dizi analiz yapılmıştır.

4.3.1. 3 sınıflı duygu analiz çalışmaları

3 Sınıflı analiz işlemleri sonucunda insanların aşı ve aşı markalarına ilişkin farklı duygu değişimleri gösterdiği gözlemlenmiştir. Bu analizler ışığında veriler üzerinde, içerisinde aşı ve aşı markaları geçen yorumlara ilişkin duygu değişimleri incelenmiştir. İncelemeler sonucunda Aşı kavramına ilişkin insanların daha olumsuz duygular içerisinde olduğu görülmüştür. Aşı markaları için ise hastalıktan kurtulma anlamında bir umut teşkil ettiği için duygu değişimlerinin daha çok olumlu yönde seyrettiği gözlemlenmiştir. Şekil 4.3.'te Aşı ve Aşı markalarına ilişkin duygu değişimlerini gösteren grafikler gösterilmiştir.



Şekil 4.3. Aşı ve Aşı Markalarına İlişkin 3 Sınıflı Duygu Değişimleri

Covid-19 virüsünün etkili olduğu ilk anlardan itibaren çekilen veriler üzerinde 2020 ve 2021 yıllarına ait insanların aylara göre farklı duygu değişimleri 2020 yılı için Şekil 4.4.'te ve 2021 yılı için ise Şekil 4.5.'te gösterilmiştir.

Şekil 4.4. ve Şekil 4.5.'de verilerin dağılımı ilgili sütunlarda farklı renklerle belirtilmiştir. Sarı olanlar ilgili sütundaki en yüksek değerleri, mor olanlar ise en

düşük değerleri gösterir. Aynı değere sahip hücreler, aynı renk kodlarıyla temsil edilir. Örneğin Şekil 4.4.'te neutral olarak belirtilen sütunda, Nisan ayında en yüksek değere sahip hücre sarı ile, Ocak ayında en düşük hücre ise mor ile işaretlenmiştir. Ayrıca Şubat, Mart, Eylül ve Ekim ayları aynı değerlere sahip oldukları için yeşil renkle gösterilmiştir.

Şekil 4.4.'e göre 2020 yılı sonlarına doğru daha olumlu yönde duygu değişimleri yaşanmıştır. 2020 yılının başından sonuna doğru ise olumsuz yönde duyguların yakın oranlarda seyrettiği, neutral olarak görülen oranların ise azaldığı görülmüştür.

	negative	neutral	positive
Ocak	39.000000	24.000000	37.000000
Şubat	40.000000	23.000000	37.000000
Mart	39.000000	23.000000	38.000000
Nisan	39.000000	25.000000	36.000000
Mayıs	39.000000	24.000000	37.000000
Haziran	39.000000	24.000000	37.000000
Temmuz	38.000000	24.000000	38.000000
Ağustos	39.000000	24.000000	37.000000
Eylül	40.000000	23.000000	37.000000
Ekim	38.000000	23.000000	39.000000
Kasım	39.000000	22.000000	39.000000
Aralık	40.000000	22.000000	38.000000

Şekil 4.4. 2020 Yılı Aylara Göre Duygu Değişimleri

2021 yılı için Şekil 4.5.'e baktığımızda 2021 sonlarına doğru 2020'nin aksine olumsuz duyguların giderek azaldığı görülmektedir. Aynı şekilde 2021 başlarından itibaren ise sonlara doğru olumlu yönde bir duygu değişiminin olduğu söylenebilir. Bu aylardaki değişimlerin aşıların bulunması ile insanların sosyal medyada bu konular hakkında daha olumlu konuşmaları gösterilebilir.

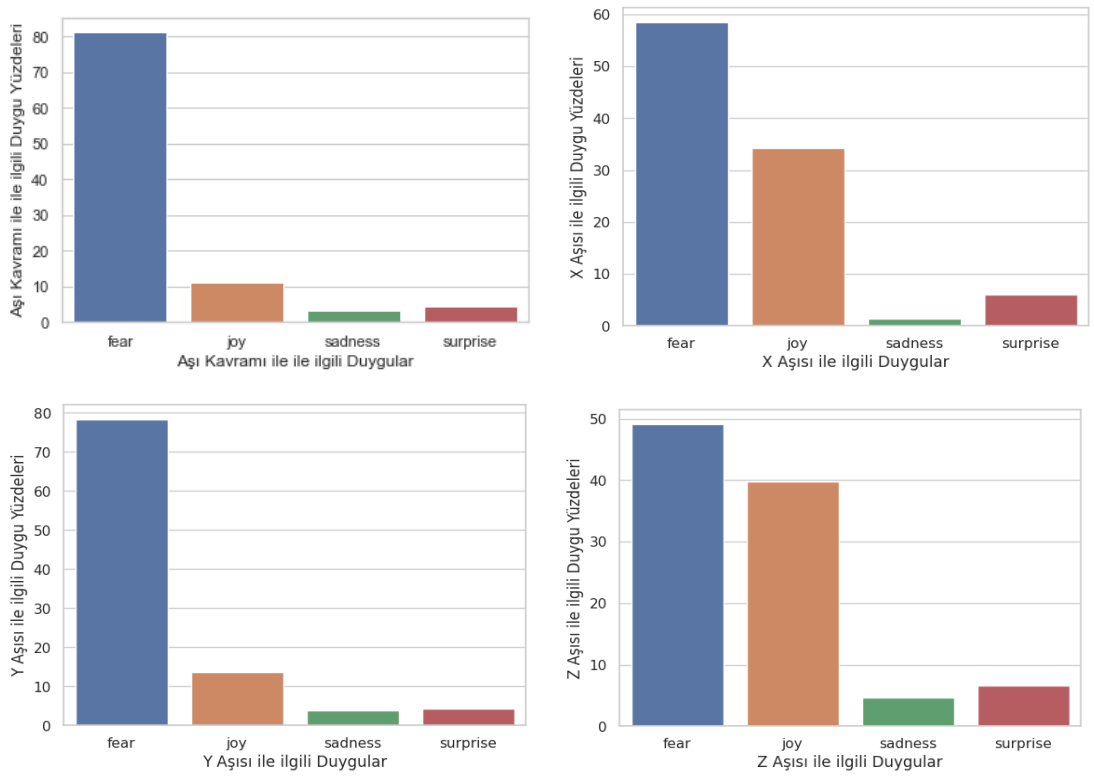
	negative	neutral	positive
Ocak	37.000000	22.000000	41.000000
Şubat	38.000000	20.000000	42.000000
Mart	38.000000	21.000000	41.000000
Nisan	36.000000	20.000000	44.000000
Mayıs	35.000000	22.000000	43.000000
Haziran	36.000000	19.000000	45.000000
Temmuz	34.000000	20.000000	46.000000
Ağustos	33.000000	19.000000	48.000000
Eylül	34.000000	19.000000	47.000000
Ekim	32.000000	20.000000	48.000000
Kasım	34.000000	18.000000	48.000000
Aralık	33.000000	18.000000	49.000000

Şekil 4.5. 2021 Yılı Aylara Göre Duygu Değişimleri

4.3.2. 4 sınıflı duygu analiz çalışmaları

4 Sınıflı duygular 3 sınıflı duygular gibi daha kutuplu bir yapıda olmadığı için farklı sonuçlar gözlemlenebilir. Örneğin fear (korku) ve sadness (mutsuzluk) duyguları olumsuz bir duygu yapısında düşünülebileceği gibi joy (Eğlenceli) de olumlu olarak nitelendirilebilir. Ayrıca surprise (şaşırtıcı) ise daha farklı bir duygu olarak karşımıza çıkmaktadır.

Şekil 4.6.'da görüldüğü gibi Aşı ve Aşı markalarına ilişkin fear (korku) duygusunun diğer duygulara daha yoğun yaşadığı görülmektedir. 3 Sınıflı duygu değişimlerinde olumlu ve olumsuz görüşlerin birbirine yakın olmasına rağmen Şekil 4.6.'da ortaya çıkan korku duygusundaki fazlalık göze çarpmaktadır. Sosyal medyada yapılan spekülasyonlar, aşıya karşı olan farklı senaryolar, virüsün oluşturduğu genel olumsuz ortam insanların genel olarak korku duymalarına neden olmaktadır.



Şekil 4.6. Aşı ve Aşı Markalarına İlişkin 4 Sınıflı Duygu Değişimleri

Covid-19 virüsünün etkili olduğu ilk anlardan itibaren çekilen veriler üzerinde 2020 ve 2021 yıllarına ait insanların aylara göre 4 Sınıflı duygu değişimleri 2020 yılı için Şekil 4.7.'de ve 2021 yılı için ise Şekil 4.8.'de gösterilmiştir.

Şekil 4.7. ve Şekil 4.8.'de verilerin dağılımı ilgili sütunlarda farklı renklerle belirtilmiştir. Sarı olanlar ilgili sütundaki en yüksek değerleri, mor olanlar ise en düşük değerleri gösterir. Aynı değere sahip hücreler, aynı renk kodlarıyla temsil edilir.

Şekil 4.7.'ye göre 2020 yılı sonlarına korku şeklinde görülen duyguların giderek azaldığı gözlemlenmiştir. 2020'nin başlarında oluşan mutsuzluk durumunda belirgin bir azalma yaşanırken aynı şekilde olumlu bir duygu olarak nitelendirebileceğimiz eğlenceli duygu durumunda olumlu yönde bir değişim görülmektedir. 2020'nin ilk aylarında ortaya çıkan şaşırtıcı duygu durumu ise virüsün hızlıca yayılması, gelen kısıtlamalar, değişen hayat biçimi gibi faktörlerin insanlarda bu yönde bir duygu oluşturmasına dayanmaktadır.

	fear	joy	sadness	surprise
Ocak	48.000000	16.000000	18.000000	18.000000
Şubat	48.000000	18.000000	18.000000	16.000000
Mart	47.000000	16.000000	16.000000	19.000000
Nisan	48.000000	19.000000	17.000000	16.000000
Mayıs	46.000000	15.000000	17.000000	22.000000
Haziran	46.000000	18.000000	15.000000	21.000000
Temmuz	45.000000	17.000000	15.000000	22.000000
Ağustos	47.000000	17.000000	14.000000	22.000000
Eylül	44.000000	19.000000	15.000000	22.000000
Ekim	45.000000	19.000000	16.000000	20.000000
Kasım	47.000000	20.000000	15.000000	18.000000
Aralık	43.000000	21.000000	16.000000	20.000000

Şekil 4.7. 2020 Yılı Aylara Göre 4 Sınıflı Duygu Değişimleri

Şekil 4.8.'de belirtildiği gibi 2021 yılına baktığımızda ise 2020 yılını devamı niteliğinde olumlu yönde seyrettiği görülmektedir. Bunun nedeninin aşının bulunması, bağışıklık kazanan insan sayısındaki artış ve günlük açıklanan virüslü sayısındaki azalış gösterilebilir. Ayrıca korku ve mutsuzluk gibi olumsuz duygulardaki azalışın bir sebebi de kısıtlamaların kaldırılması, eski sosyal hayata geri dönüş gibi nedenler sayılabilir.

	fear	joy	sadness	surprise
Ocak	48.000000	16.000000	18.000000	18.000000
Şubat	48.000000	18.000000	18.000000	16.000000
Mart	47.000000	16.000000	16.000000	19.000000
Nisan	48.000000	19.000000	17.000000	16.000000
Mayıs	46.000000	15.000000	17.000000	22.000000
Haziran	46.000000	18.000000	15.000000	21.000000
Temmuz	45.000000	17.000000	15.000000	22.000000
Ağustos	47.000000	17.000000	14.000000	22.000000
Eylül	44.000000	19.000000	15.000000	22.000000
Ekim	45.000000	19.000000	16.000000	20.000000
Kasım	47.000000	20.000000	15.000000	18.000000
Aralık	43.000000	21.000000	16.000000	20.000000

Şekil 4.8. 2021 Yılı Aylara Göre 4 Sınıflı Duygu Değişimleri

Yukarıdaki analiz sonuçlarına göre insanlarda aşı ile ilgili oluşan duyguların aslında hayatın tüm faktörleri ile birlikte değişim gösterdiği söylenebilir. Yani bir insan aşı ve aşı markalarına çok olumlu duygular içerisinde olsa da hayatını yaşamasına engel faktörler genel ruh halini olumsuz yönde etkileyebilir.

5. SONUÇLAR ve ÖNERİLER

Yapılan çalışmada Sosyal Medyadan Koronavirüs salgını ve Aşı ile ilgili kullanıcı yorumları çekilerek kullanıcıların duygu durumlarını ortaya çıkarmaya çalışılmıştır. Reddit Sosyal Medya Platformundan Koronavirüs ile ilgili konulardan 2020 ve 2021 yılına ait veriler aylık olarak düzenli bir formda çekilmiştir. Farklı alanlarda da veri çekme işlemlerine katkı sağlaması için Python dilinde DataExtract isminde bir sınıf oluşturulmuştur. Bu sınıf sadece ham verileri çekmekle kalmayıp aynı zamanda veri temizleme işlemini de otomatik olarak yapmaktadır. Ayrıca kullanıcılar veriler çekilirken farklı json dosyalarında tutulmuştur. Daha sonra etiketsiz olarak çekilen veriler üzerinde duygu durumlarını ortaya çıkarmak için çok sözlük tabanlı 2 farklı modelde duygu sınıflandırıcıları kullanılmıştır. Vader ile veriler 3 sınıflı olarak pozitif, negatif ve nötr olarak etiketlenmiştir. Daha sonra aynı veriler etiketsiz olarak sparkNLP ile 4 sınıflı korku, mutsuzluk, eğlenceli ve şaşırtıcı olarak etiketlenmiştir. Reddit'te bulunan yorumlarda karakter sınırı olmaması kullanıcıların genel bir duygu durumlarının ortaya çıkmasını zorlaştırmıştır. Bu sebeple etiketlenmiş veriler bir dizi eğitim işlemlerine tabi tutulmuştur. Glove kelime gömme yapısı kullanılarak veriler üzerinde eğitim işlemleri yapılmıştır. Daha sonra Bert ile tekrardan veriler eğitime tabi tutulmuştur. Vader ile 3 sınıflı veriler üzerinde 0.70 bulunan bir doğruluk oranı Glove ile 0.74 ve daha sonra Bert ile 0.82 başarım oranına ulaşmıştır. spakNLP yapısı ile 4 sınıflı veriler üzerinde başlangıçta 0.71 bulunan doğruluk oranı Glove ile 0.73 e ve Bert ile 0.78 başarım oranına ulaşmıştır.

Daha yüksek başarıma oranlarına ulaşan veriler üzerinde çalışmanın odaklandığı ana konu olan Aşı analizi işlemleri gerçekleştirilmiştir. 2020 ve 2021 verileri üzerinde ayrı ayrı gerçekleştirilen analiz sonuçları hem aylara hem de yıllara göre değişimlerin görülmesine olanak tanımıştır. Salgının başından 2021 yılının sonuna kadar geçen sürede insanların aşı bulunması sonrasında daha olumlu bir duygu içerisine girdikleri görülmüştür.

KAYNAKLAR

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.

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.

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

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.

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.

DEVLIN, J., CHANG, M.-W., LEE, K., and TOUTONOVA, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

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

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

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.

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. social. gatech. edu/papers/icwsm14. vader. hutto. pdf', p. 82.

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

HUTTO, C. and GILBERT, E. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media. 8, 1 (May 2014), 216-225.

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.

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

KIRITCHENKO, S., ZHU, X., CHERRY, C., and MOHAMMED, 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.
MANNING, C. D., RAGHAVAN, P. and SCHUTZE, H., 2009. “An Introduction to Information Retrieval” online, Retrieved from the internet.
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, 675-679.
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.
PANG, BO, LEE, L. and VAITHYANATHAN, A., 2002. "Thumbs up? Sentiment classification using machine learning techniques." arXiv preprint cs/0205070.
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.
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.
SINGHAL, T., 2020. A Review of Coronavirus Disease-2019 (COVID-19). Indian J Pediatr 87, 281–286.
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.
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. semi-supervised sentiment classification." Neurocomputing 131: 312-322.
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.
WEIDMAN, A. C., STECKLER, C. M., and 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.
YIN, D., MENG, T. and CHANG, K.-W., 2020. ‘Sentibert: A transferable transformer-based architecture for compositional sentiment semantics’, *arXiv preprint arXiv:2005.04114* .
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.
YUWEN, Z., ZHAOZHUO, X., 2018. Bert for question answering on squad 2.0. Unpublished manuscript.
ZHOU, SHUSEN, CHEN, Q. and WANG, X., 2014. "Fuzzy deep belief networks for question answering on squad 2.0,Unpublished manuscript.