ANKARA ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

BLM 462 PROJE RAPORU

Makine Öğrenmesi Yöntemleriyle Sosyal Medya Analizi

Emir Utku BİCAN

16290012

Doç. Dr. Semra GÜNDÜÇ

Mayıs, 2021

ÖZET

Bu rapor 2021 Bahar yarıyılında tamamlanan Makine Öğrenmesi Yöntemleri ile Sosyal Medya Analizi bitirme projesi için hazırlanmıştır. Projede Logistic Regression ve Stochastic Gradient Descent kullanılarak bir duygu sınıflandırması modelinin nasıl geliştirildiği ve bu model kullanılarak elde edilen analiz sonuçları sunulmuştur. Ayrıca duygu sınıflandırma problemine sunulan bir çözüm olmasının yanı sıra belirlenen periyotta toplanan verilerden hareketle kullanıcıların fikir ve alışkanlarını keşfetmek amaçlanmış ve bu analizin sonuçları da sunulmuştur.

İÇİNDEKİLER

1. GİRİŞ

İnternet kullanımının her geçen gün artmasıyla sağlanılabilecek yararlar da çeşitlenmektedir. Bu artış sayesinde kullanıcıların sosyal medya aracılığıyla ürettiği içerik, eldeki verinin de hacminin artmasına önayak olmaktadır. Böyle bir verinin firmalar, hükümetler ve toplum için içgörü sağlaması bu yararların arasındaki en kritik örneklerdendir.

Sosyal medya analizi, şirketlerin verilerden anlamlı kalıplar ve eğilimler belirlemesini sağlayarak, şirketlerin ürün geliştirmeden halkla ilişkilere kadar değişen alanlarda bilinçli kararlar almasına yardımcı olur. Bu kararlar alınırken, şirketlerin müşterilerinin ihtiyaçlarını ve beklentilerini daha iyi anlamasına, sosyal kanallarda yürütülen müşteri hizmetleri ve pazar araştırmalarının verimliliğini arttırmasına, ürün geliştirme ve pazarlamaya daha akıllıca yatırım yapmasına ve rekabet akıllarını arttırmasına yardımcı olabilir.

Markalar sosyal medya kullanımının kullanıcılar arasında popülerlik kazanmasıyla reklam politikalarını buna adapte etmeye başlamışlardır. Bu adaptasyon çalışmalarının bir getirisi olarak çok takipçili kullanıcıların etki kuvveti sebebiyle sosyal medya fenomenliği(influencer) adında yeni bir iş kolu ortaya çıkmıştır. Mediakix isimli influencer pazarlama ajansı tarafından yapılan bir analize(2021) göre halihazırda TikTok, Youtube ve Instagram platformları üzerinde influencerlik yapan kişi sayısı tahmini 3.2 - 37.8 milyondur. Influencer ekosisteminin sürekli hacim kazanması dolayısıyla sosyal medya analizi de önem kazanmıştır. Kullanıcılar kitlelerinin eğilimlerini anlamak ve bundan hareketle içeriklerini şekillendirmek amacıyla bahsedilen platformların kendilerine sağladığı analiz yöntemlerini kullanmakta ve rakipleriyle kıyas yaparak stratejiler belirleyebilmektedir.

Genel kapsamda düşünülecek olursa, influencer veya kullanıcı etkileşimi ile açığa çıkan her içerik kamuoyunun görüşü hakkında bilgiler sunmaktadır. Bu projede çalışılacak problem de buradan ortaya çıkmıştır. 2019 Kışı itibariyle dünyayı etkisi altına alan pandemi aynı anda sosyal medyayı da etkisi altına almıştır. Ve bu etki insanların duygularını anlatmaları için etkili bir yöntem olan yazıyı temel alan Twitter'den toplanan veriler üzerinden incelenebilir. Twitter, kullanıcıların fikirlerini

tweetlere dökerek oluşturduğu popüler bir mikroblog hizmetidir. Bu tweetler öyle ya da böyle konuşulan konular hakkında duygu belirtir. Bu duygunun anlamlandırılması için pek çok analiz projesi yapılmaktadır. Bu projede de pandemi mücadelesinde, aşılar hakkında atılan tweetlerin gözlemlenmesi amaçlanmıştır. (Go vd., 2009)

1.1. Veri Kümesi

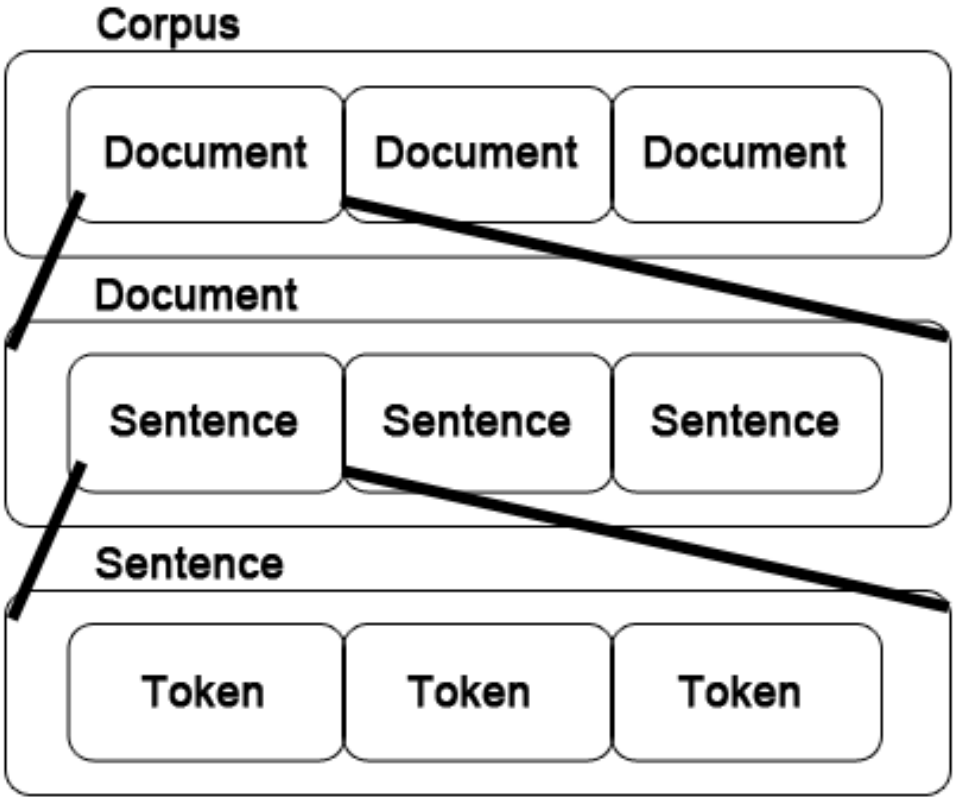
Proje genel hatlarıyla veri toplama, veri ön işleme ve analiz olmak üzere 3 aşamadan oluşmaktadır. İlk aşamanın sonunda pandemi ile ilgili anahtar kelimeler ve aşılar hakkında Twitter'den çekilen verilerin bulunduğu birincil veri kümesi oluşturulmuştur. Birincil veri kümesi toplanan tweetlerin metinlerine ek olarak analiz edilmesi için farklı değerler de içermektedir.

İkincil veri kümesi olarak, Kaggle sitesi üzerinde paylaşılmış ve belirli dönemler içerisinde güncellenen bir veri kümesi kullanılmıştır. Bu veri kümesinde aşılar hakkında daha özel tweetler bulunmaktadır. Kaggle veri kümesi kullanılarak 1340 adet elle etiketleme işlemi gerçekleştirilmiştir.

Birincil ve ikincil veri kümeleri birleştirilerek birleşik veri kümesi oluşturulmuştur. Birleşik veri kümesi model eğitimi için kullanılacaktır. Devamında Kaggle veri kümesi üzerinde eğitilen model kullanılarak duygu sınıflandırması yapılacaktır.

1.2. Doğal Dil İşleme ve Duygu Analizi

İnsanlık ve dil paralel olarak gelişen iki kavramdır. Tarihin var olmasından bu yana insanlar etkileşimde bulunmuş ve bunun için dil adında bir yöntem geliştirmiştir. "Bilişim teknolojilerindeki gelişmeler, bilgisayarlı dil bilimi çalışmalarına önemli bir ivme kazandırmıştır. Doğal Dil İşleme (DDİ) adı verilen bu yeni bilim alanı önceleri insan bilgisayar etkileşiminde doğal dillerin kullanılabilmesi amacıyla başlatılmış, zamanla bilgisayarlı dil bilimine dönüşmüştür." Doğal dil işleme, kelime kökleri ve metindeki dizilişleri ile konuşma bağlamından hareketle, dilin bilgisayar tarafından anlaşılabilir hale getirilmesi işlemlerinden başlayarak yazılı metnin okunması, özetinin çıkarılması, içeriğinin anlaşılması ve bilgiye dönüştürülmesi, chatbotlar ve metin çevirisi gibi özel problemlere yoğunlaşır. (Adalı, 2016)



Şekil 1.1. Doğal dil işleme hiyerarşisi

Doğal dil işlemenin çalışma konularından birisi olan duygu analizi için sözlük tabanlı veya makine öğrenimi tabanlı yaklaşımlar kullanılmaktadır. Sözlük tabanlı yaklaşımda metnin kutupsallığı ve öznelliği, kelimelerin duygusal değerlerini içeren bir sözlükten hareketle hesaplanır. Makine öğrenmesi temelli yaklaşımlarda ise metinlerin sayısal değerlere dönüştürülerek, külliyatta(corpus) bulunan verinin dağılımlarından hareketle çeşitli makine öğrenmesi algoritmalarının uygulanarak sınıflandırma veya kümeleme yapılması söz konusudur.(Kharde vd., 2016)

2. MATERYAL VE YÖNTEM

2.1. Verilerin Toplanması

Twitter, geliştirici hesap başvurusu yapan kullanıcılarına sınırlı sayıda erişim sunan bir API bulundurmaktadır. Proje için gerekli veri bu API yardımıyla 15 Mart-15 Mayıs günleri arasında; vaccine, corona, coronavirus, lockdown, covid, covid19, biontech, astrazeneca, sinovac, sputnik, moderna anahtar kelimeleri kullanılarak belirli dönemler içerisinde anlık akıştan toplanan tweetleri içerir. Metin verisine ek olarak; kullanıcı adı, tweetin atıldığı konum ve tarih, kullanıcının takipçi sayısı, kullanıcının takip ettiği kişi sayısı ve hesabın oluşturulduğu tarih, tweetlere yapılan retweet sayıları ile metin içerisinde geçen etiketler de toplanmıştır. Twitter API tarafından belirlenen kotanın dolması sebebiyle mayıs ayı itibariyle veri toplama işlemine son verilmiştir. Buna ek olarak API yardımıyla toplanan veri kümesi, Kaggle üzerinde paylaşılan, Gabriel Preda tarafından oluşturulmuş ve düzenli olarak güncellenen yine bahsedilen içeriklerin bulunduğu *'All COVID19 Vaccines Tweets'* veri kümesi ile harmanlanmıştır.

Tweetler aşağıda sözde kodu belirtilen metot ile toplanmıştır. Metot çalıştırılmadan önce iterasyon sayısı(run) = 6, toplanacak tweet sayısı(count) = 2500 olarak belirlenmiş ve her iterasyondan sonra 15 dakikalık bekleme ile Twitter API kısıtlarına uygun olarak aramalar yapılmıştır.

```
Giriş: run, count, keywords, since, language
Çıkış: res
Scraping:
     Search(run, count, keywords, since, language)
          res += username
          res += location
          res += following
          res += followers
          res += date_acc_create
          res += date_tweet_create
          res += text
          res += hashtag
     Return res
     Cooldown 15 mins
```

Belirtilen yöntem yardımıyla, haftalık ortalama 17000 tweet toplanmıştır. Bu işlemlerden sonra verilerin analize hazırlanması için metin temizleme ve metin etiketleme işlemleri uygulanmıştır.

2.2. Verinin Ön İşlenmesi

Toplanan verinin bilgisayar tarafından anlamlandırılması için metin temizleme ve etiketleme ön işlemleri yapılması gerekliliği doğmuştur.

2.2.1. Metin Temizleme

Toplanan metinlerin analiz edilmek üzere hazırlanması için yapılan işlemlerdir. Metnin bilgisayar tarafından anlamsal olarak anlamlandırılmasını güçleştiren linkler, hashtaglar, noktalama işaretleri, rakamlar ve anlamda değişiklik yapmayan kelimeler(stopwords) bu aşamada temizlenmiştir. Metin üzerinde bütünlük elde etmek için metinlerdeki büyük-küçük harf ayrımı ortadan kaldırılmıştır. Metin en basit hal olarak kelimelerine parçalanmıştır. Bu anlatılan işlemlerden sonra, metnin bilgisayar tarafından anlaşılır hale getirilmesi için sayısal olarak temsil edilmesi gerekmektedir. Bu dönüşümün verimli bir şekilde yapılması için metinlerin elemanlar en yalın hallerine indirilmişti. Sonuca ulaşmak için stemming ve lemmatization olmak üzere iki işlem bulunmaktadır. Stemming kelimenin kökünün morfolojik varyantlarının bulunması, lemmatization ise kelime gövdesinin bulunması ve buna ek olarak bağlamdaki kelimelerin tek bir kelimeye bağlanmasıdır.

Çizelge 2.1. Stemming ve lemmatization örnekleri

Stemming Örnekleri	**Lemmatization Örnekleri**
retrieval/retrieved/retrieves -> retrieve	rocks -> rock
likes/liked/likely/liking -> like	better -> good

Projede lemmatization kullanılmasının ana sebebi WordNet külliyatı kullanarak stopwordleri elemesi ve kelimeleri türlerine göre etiketleyip dönüştürmesi dolayısıyla yavaş çalışmasına rağmen, sonuçta elde edilen tokenlerin yine gerçek kelimeler olmasıdır. Stemming işleminde sonuçta elde kalan token çoğu zaman anlamsız kelimeler olmaktadır.(happiest->happi, chocolate->choco vb.) NLTK kütüphanesinin WordNetLemmatizer algoritması yardımıyla bu işlem gerçekleştirilmiştir.

2.2.2. Metin Etiketleme

'All COVID19 Vaccines Tweets' veri kümesinde, resim, ses, metin ve zaman serileri gibi birçok veri türünü etiketlemek ve keşfetmek için kullanılan açık kaynaklı veri etiketleme aracı Label Studio kullanılarak elle etiketleme yapılmıştır.

Metinler etiketlenirken aynı zamanda verinin elenmesi işlemi de söz konusu olmuştur. İşlem boyunca karşılaşılan örnekler, olumlu bir söz diziminde olmasına rağmen olumsuz anlama gelmesi, mecaz içermesi durumunda veya aynı metin içerisinde birden fazla aşının karşılaştırılması halinde etiketleme yapılmadan elenmiş ve bunun sonrasında 1340 adet etiketlenmiş veri elde edilmiştir.

Çizelge 2.2. Elle etiketleme ve eleme işleminde yapılan bazı elemeler ve sebepleri

ID	Metin	Sebep
1	I heard the side effects of moderna are worse than Pfizer. Good luck.	Karşılaştırma
2	You can't. Sinovac is whole virus vaccine. Pfizer is mrna. Sabi, one may combine only when it is of the same activation type (like pfizer and modern which both use mrna)	Karşılaştırma
3	I didn't say we were. My point is, the Chinese Sinovac vaccine is nowhere near as effective as the Pfizer and OxfordAZ vaccines	Karşılaştırma
4	I didn't say we were. My point is, the Chinese Sinovac vaccine is nowhere near as effective as the Pfizer and OxfordAZ vaccines	Karşılaştırma
5	just got the sinovac jab, so when am i gonna turn into a titan	Mecaz

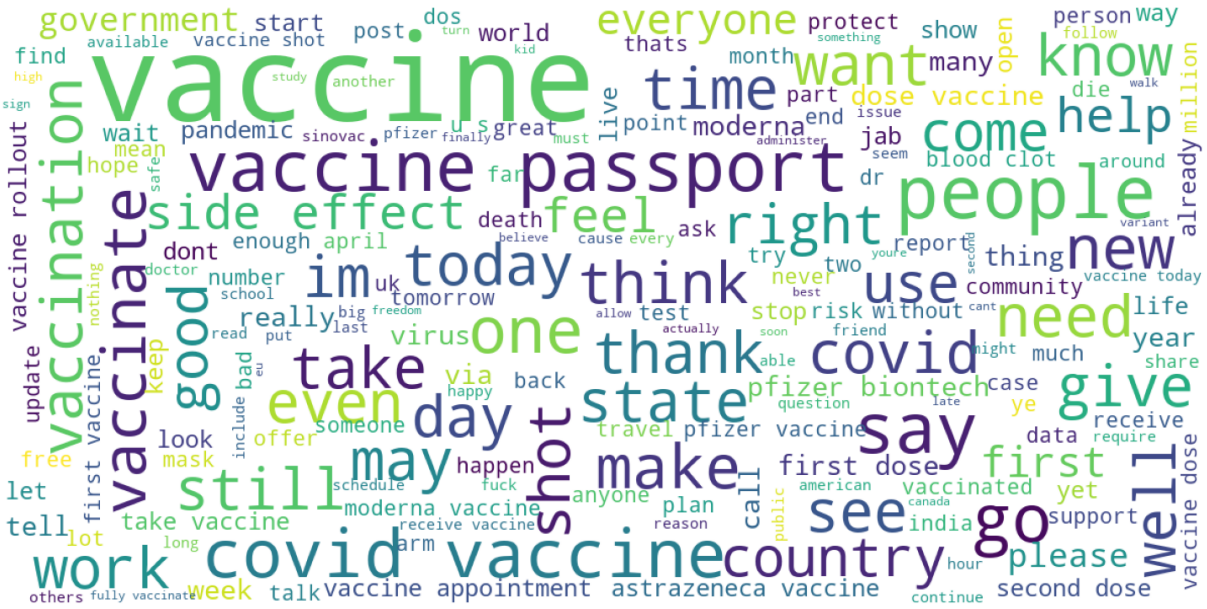
Bu projede elle etiketleme işlemine ek olarak birincil veri kümesi de, NLTK kütüphanesinde tanımlı VADER duygu analiz modeli kullanılarak etiketlenmiştir. VADER, kelime özellikleri ile kutupsallık puanlarını eşleştiren sözlük tabanlı duygu analizi modelidir. Bu modelde metin verisinin duygu puanı, metindeki her bir kelimenin kutupsallık puanlarının toplanmasıyla elde edilmektedir. Eldeki veriye uygulanması sonrasında aşağıdaki formüle göre etiketleme tamamlanmıştır.

$$sentiment = \begin{cases} Olumlu, & 1.0 \geq Duygu\ Puanı \geq 0.1 \\ Nötr, & 0.1 > Duygu\ Puanı > -0.1 \\ Olumsuz, & -0.1 \geq\ Duygu\ Puanı \geq -1.0 \end{cases}$$

Yapılan bu işlemlerden sonra elde edilen hibrit veri kümesi kullanılarak, makine öğrenmesi modeli geliştirilmiş, sonrasında sadece aşılarla ilgili tweetlerin bulunduğu veri kümesi üzerinde duygu sınıflandırması yapılmıştır.

2.3. Makine Öğrenmesi Algoritmaları ile Duygu Sınıflandırma

Açık kaynaklı makine öğrenmesi kütüphanesi scikit-learn yardımıyla, Logistic Regression ve Stochastic Gradient Descent algoritmaları OneVsRestClassifier kullanılmış ve duygu sınıflandırma modelleri geliştirilmiştir.



Şekil 2.1. Model eğitim külliyatı(Training Corpus) kelime bulutu

Şekil 2.1.'de görülen kelime bulutu, duygu sınıflandırma modeli geliştirilirken eğitim verisi olarak kullanılan veri kümesinden hareketle oluşturulmuştur. Elde edilen bu küme TF-IDF yöntemi ile sayısallaştırılmış ve N-Gram yöntemi kullanılarak model eğitilmiştir.

2.3.1. Terim Sıklığı – Ters Doküman Sıklığı (TF-IDF) ve N-Gram Modeli

TF-IDF cümlede bulunan bir kelimenin genelle ne kadar alakalı olduğunu değerlendiren bir istatistiksel ölçü yöntemidir. Terim sıklığı ve ters doküman sıklığı metriklerinin birbiriyle çarpılmasıyla elde edilir. Bir cümlede veya veri kümesinin tamamında kelimenin ne kadar tekrar ettiği sayılır ve bu kelimenin bulunduğu belgenin tekrarıyla çarpılarak dengelenir.

Terim sıklığı, bir kelimenin tüm veri kümesi içinde ne kadar tekrar ettiğini sayarken bu kelimenin hangi sınıfta nasıl bir yoğunlukta olduğunun da hesaplanması ile elde edilir.

$$tf(t,d) = \log\big(1 + freq(t,d)\big)$$

$$idf(t,D) = \log\left(\frac{N}{count(d \in D : t \in d}\right)$$

$$tfid(t,d,D) = tf(t,d).idf(t,D)$$

Makine öğrenmesinin temeli sayılar olması sebebiyle metin vektörizasyonu bu yöntemle yapılır. Benzer kelimelere sahip cümlelerin benzer vektörleri olur ve bu sayede korelasyon keşfi yapılabilir.

Doğal dil işleme ve olasılık alanlarında, bir n-gram, belirli bir metin veya konuşma örneğinden oluşturulan n elemanlı oluşan bitişik bir dizidir. Bir elemandan sonra gelen öğenin tahmin edilmesi için kullanılan bir yöntemdir.(Broder vd., 1997)

2.3.2. OneVsRest Çatı Modeli

OneVsRest ikili sınıflandırma algoritmalarının çoklu sınıflandırma yapılması için kullanılmasını sağlayan bir yöntemdir. İki sınıflı olmayan veri kümelerinin birden çok ikili sınıflandırma problemine bölünmesini sağlar. Örneğin, "kırmızı", "mavi" ve "yeşil" için örnekler içeren çok sınıflı bir sınıflandırma problemi verildiğinde OneVsRest Classifier kullanılarak problem, gibi ikili sınıflandırma problemlerine ayrılır.(Brownlee, 2020)

- İkili Sınıflandırma Problemi 1: kırmızı veya [mavi veya yeşil]
- İkili Sınıflandırma Problemi 2: mavi veya [kırmızı veya yeşil]
- İkili Sınıflandırma Problemi 3: yeşil veya [kırmızı veya mavi]

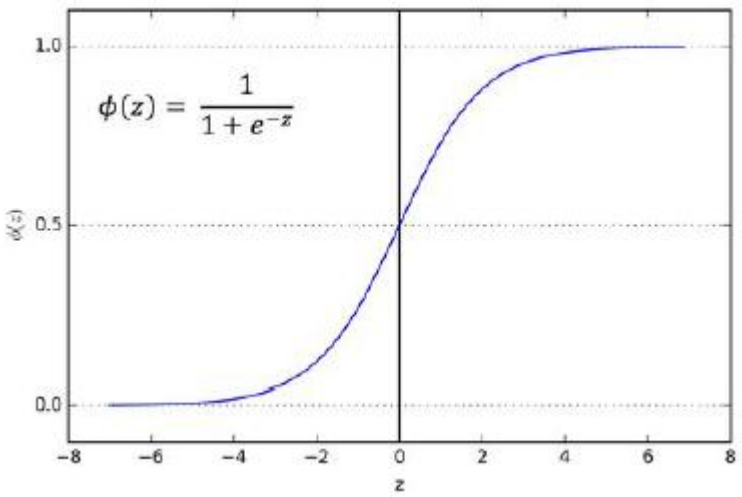
Duygu sınıflandırması probleminde de duyguların pozitif, nötr ve negatif olmak üzere üç sınıf olarak ifade edilmesi sebebiyle bu yöntem kullanılmıştır.

2.3.3. Logistic Regression

İki olası sonucu olan bir olayı veya sınıfı modellemek için lojistik bir işlev kullanılan, ikili regresyon formunun parametrelerini tahmin edilmesini sağlayan istatistiksel bir modeldir. Doğrusal çıktının hesaplanmasının ardından regresyon çıktısı üzerinden bir saklama işlevi izler. Bu saklama işlevinde en çok kullanılan yöntem sigmoid fonksiyonudur.

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots.$$
$$h(\theta) = g(z)$$
$$g(z) = \frac{1}{1 + e^{-z}}$$

Sınıflandırıcı, giriş boyutunu sigmoid fonksiyonu kullanarak bir alandaki tüm noktalar aynı sınıfa karşılık gelecek şekilde iki boşluğa böler.



Şekil 2.2. Sigmoid fonksiyonu

2.3.4. Stochastic Gradient Descent

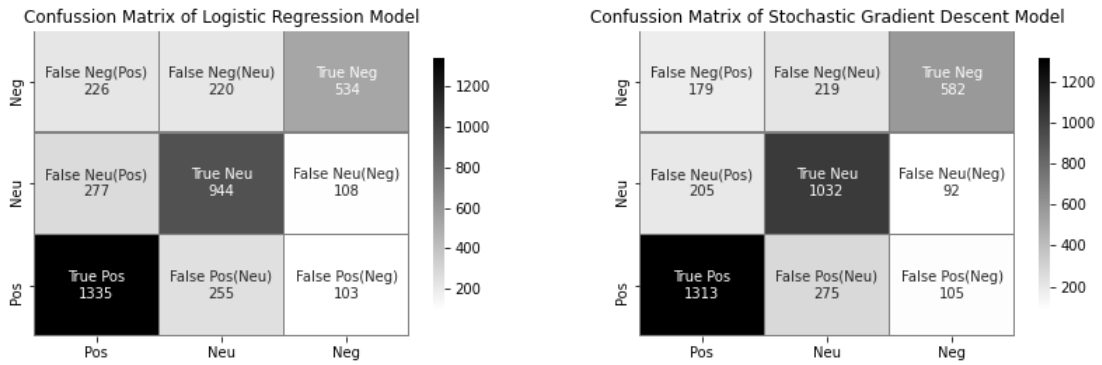
Stochastic gradient descent, uygun düzgünlük özellikleriyle bir hedef işlevi optimize etmek için kullanılan yinelemeli bir yöntemdir. Tüm veri kümesinden hesaplanan gerçek gradyanı, bunun verilerin rastgele seçilen bir alt kümesinden hesaplanan bir tahminiyle değiştirdiği için, gradyan iniş optimizasyonunun stokastik bir yaklaşımı olarak kabul edilebilir. Özellikle yüksek boyutlu optimizasyon problemlerinde bu, hesaplama yükünü azaltır ve daha düşük bir yakınsama oranı için daha hızlı yinelemeler sağlar.(Nowozin vd., 2012)

Belirlenen sayıda yineleme boyunca formüle göre sonucun hesaplanmasına yarayan bir algoritmadır.

```
Başlangıç vektörü olarak ω ve η parametrelerini seç
Tahmini minimum elde edilene kadar tekrarla:
     Eğitim setindeki örnekleri karıştır
     ω := ω − η/n ∑_{i=1}^{n} ∇Q_i(ω)

     η: Öğrenme oranı(yineleme sayısı)
```

Bu iki algoritma kullanılarak oluşturulan metin sınfılandırma modellerinden Logistic Regression modelinin başarı oranı %70,28, SGD modelinin başarı oranı %73,13 olarak elde edilmiştir. Bu sonuçlardan hareketle projenin geri kalanında kullanmak üzere SGD modeli seçilmiştir.



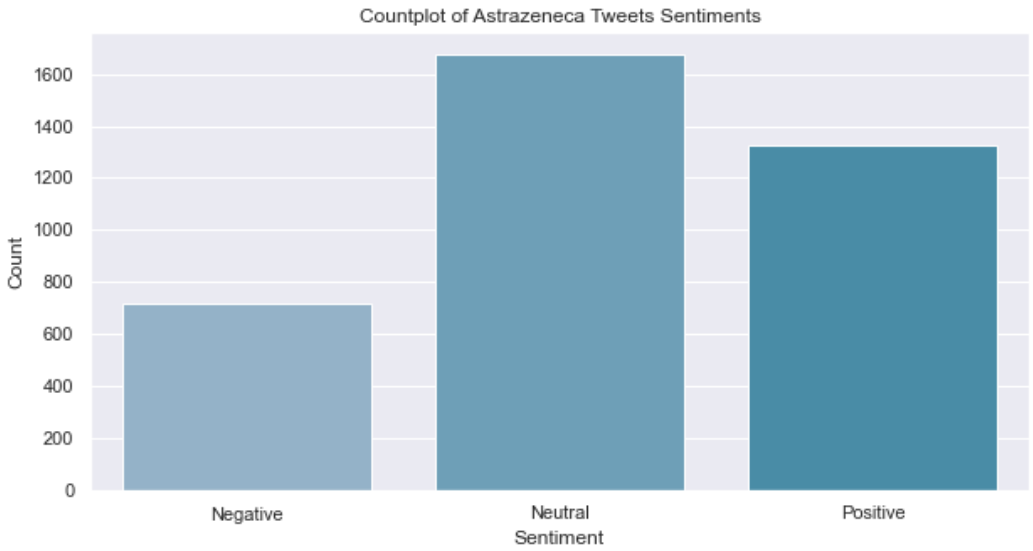
Şekil 2.3. Modellerin karmaşıklık matrisleri

3. SONUÇ

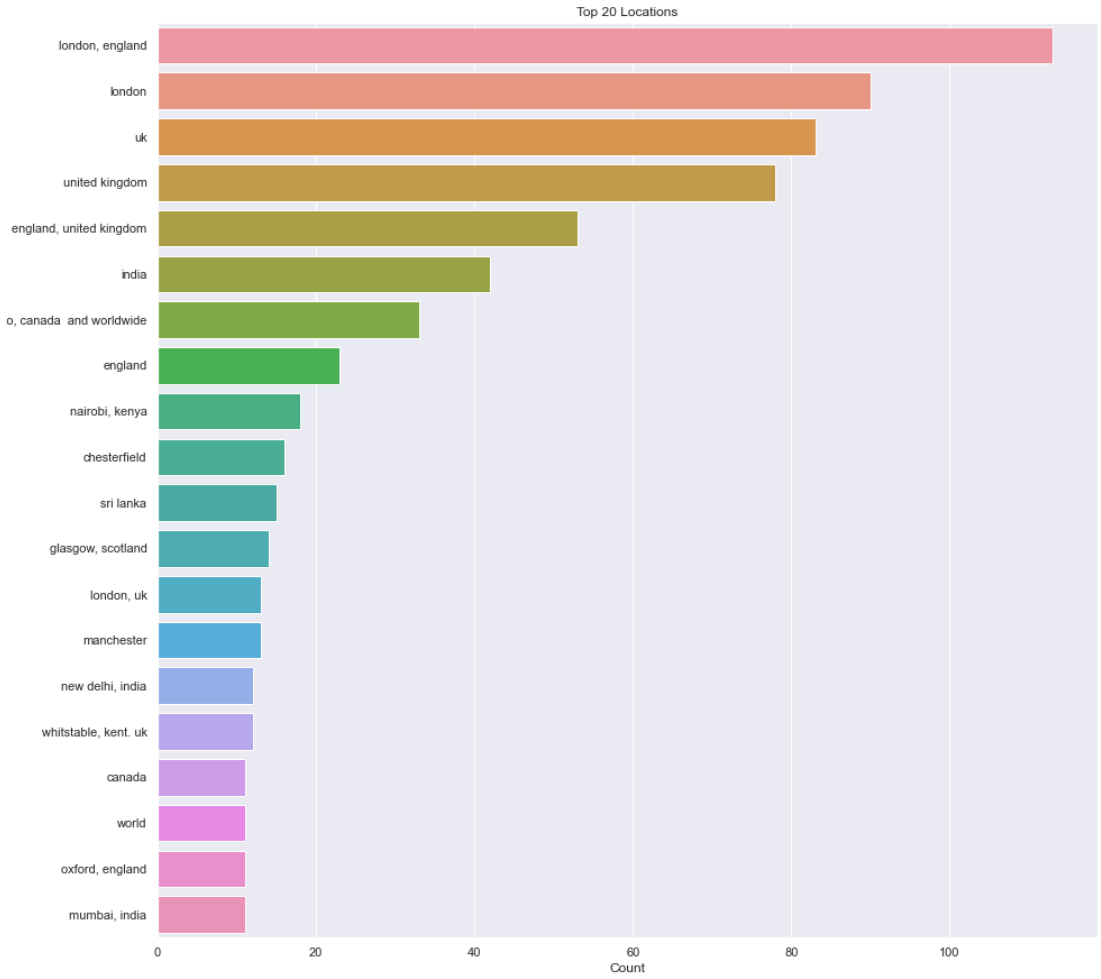
Pandemiye karşı sürdürülen mücadelenin en önemli silahı olan aşılar için Twitter'den toplanan veriler yardımıyla insanların neler düşündüğüne ışık tutmanın amaçlandığı bu projede, iki farklı etiketleme yöntemiyle elde edilen hibrit veri kümesi kullanılarak eğitilen SGD duygu sınıflandırma modelinden hareketle şu sonuçlar elde edilmiştir.

Çizelge 3.1. Örnek sınıflandırma sonuçları

text	sentimen
there is no difference btwn #astrazeneca & #pfizer vaccines given the r	Negative
if the eu don't want to use the #oxfordastrazeneca vaccine then send it	Negative
india delays big exports of #astrazeneca shot as infections surge#astr	Negative
russia's #covid19 vaccine #sputnikv approved for emergency use in #v	Negative
the government of #ukraine paid $54 million to #india for #covishieldva	Negative
did my part for humanity yesterday and got my 1st moderna vaccine.	Negative
#breaking: china admits its vaccines' effectiveness is low#breakingnev	Negative
big promises, few doses: why #russia's struggling to make #sputnikv d	Negative
@jon_christian they can have mine□am waiting until i can have the #spu	Negative
@bbchughpym your piece on @bbcnews just now failed to mention the	Negative
first moderna shot in! #moderna #vaccinated #covid19vaccine #vaccin	Neutral
#sputnikv effectiveness is 97.6%, according to the data analysis of 3.8	Neutral
booked my first vaccine shot. i'm a #moderna	Neutral
as a wonderful end to my day volunteering at a #covid vaccine site yest	Neutral
day one after getting the second #moderna vaccine - a lot of fatigue. a	Neutral
@pfizerbiontech you should be ashamed for pulling out of the sovereign	Neutral
shot #2 has been a little tricky. thankfully the side effects don't last long.	Neutral
thumbs up for #moderna #modernavaccine @billgates @elonmusk #va	Neutral
#pfizervaccine #oxfordastrazeneca #vaccinesideeffects they've adopte	Neutral
second shot in the books. two weeks until full vaccination. this is the wa	Neutral
a woman from long island tested positive for #covid19 after having take	Positive
good morning. in print today: #sinopharm, #sinovac effective, safe: #wh	Positive
@dhpsp i can't answer. i agree to receive the administration of the chin	Positive
brazil health ministry says in negotiations for moderna to deliver 13 mln	Positive
#russia's #sputnikv vaccine takes key step to #eu approval - france 24	Positive
received 2nd dose today. thank you, @nyulangonebk !! □□□#pfizerbi	Positive
had second moderna #vaccine this morning. it has been 6 hours. so fa	Positive
shot #2 down and a nice hike to boot. #moderna #covidvacccine	Positive
australia's vaccine approval:#astrazeneca #pfizerbiontech□nd now, #m	Positive



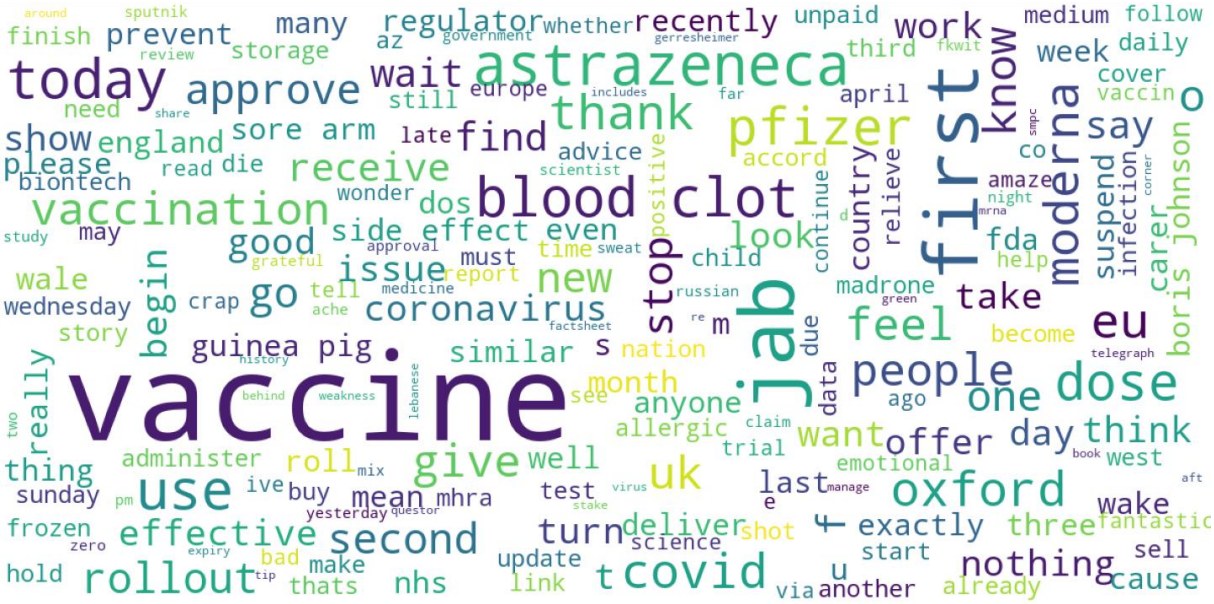
Şekil 3.1. Astrazeneca hakkında atılan tweetlerin duygu dağılımları



Şekil 3.2. Astrazeneca tweetlerinin konumları

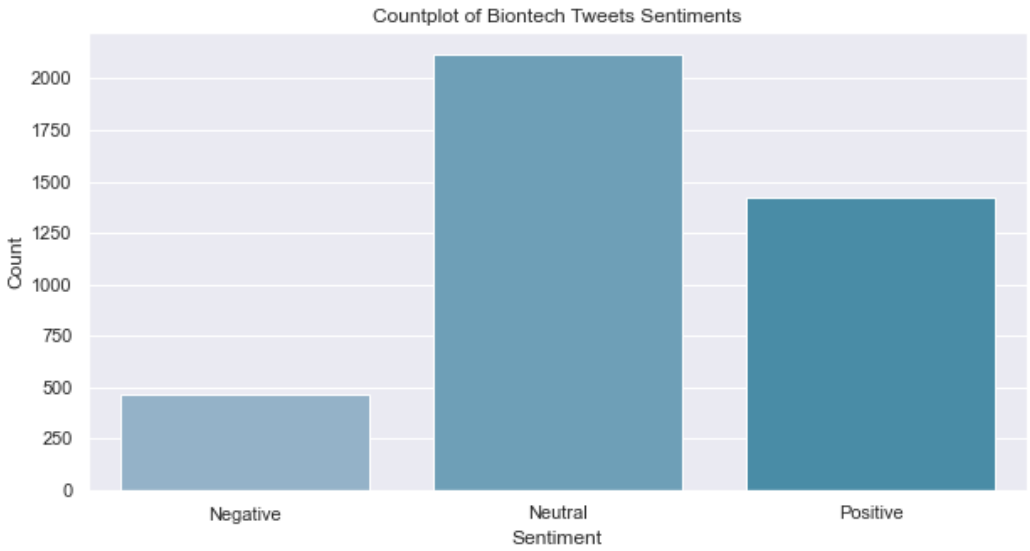
Şekil 3.1.'de görülen sütun grafiği Astrazeneca hakkında atılan olumsuz tweetlerin kıyasla daha az olduğunu göstermektedir. Buradan hareketle Astrazeneca aşısı olan Twitter kullanıcılarının genel olarak sonuçlardan memnun görülmektedir. Veriler göstermiştir ki, Astrazeneca hakkında atılan tweetlerin 737 tanesi olumsuz, 1310 tanesi olumludur. Bunlara ek olarak 1666 adet tweet de nötr olarak değerlendirilmiştir.

Şekil 3.2. ise bu tweetlerin en çok hangi konumlardan atıldığını keşfetmek için kullanılabilir. Astrazeneca aşısının Britanya-İsveç üretimi olması ve İngiltere'de kullanılması sebebiyle tweetlerin İngiltere ve çevresinde yoğunlaşmış olduğu görülmektedir.

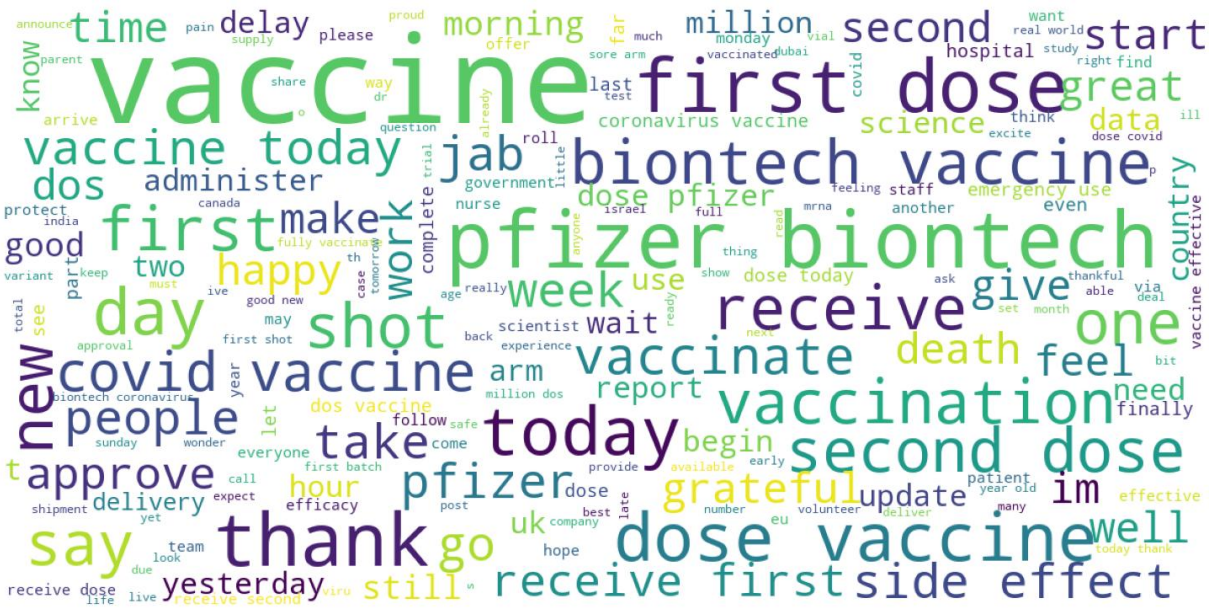


Şekil 3.3. İngiltere konumlu tweetlerden oluşturulan kelime bulutu

İngiltere konumlu tweetlere bakıldığında *Astrazeneca* göze çarpan kelimelerden biridir. Önceden de belirtildiği gibi şartlar dolayısıyla tweetlerin bu konumda yoğunlaştığı görülmektedir. Bir başka örnek olarak *blood clot* kelime grubu yoğun kullanılan kelimelerdendir. Araştırmalar yapıldığında Astrazeneca ve J&J aşıları hakkında ölümcül kan pıhtılaşmasına sebep olduğu ile ilgili haberlere erişilmiştir. Kelime bulutu sayesinde bu aşı hakkında endişelerin olduğu keşfedilmiştir.



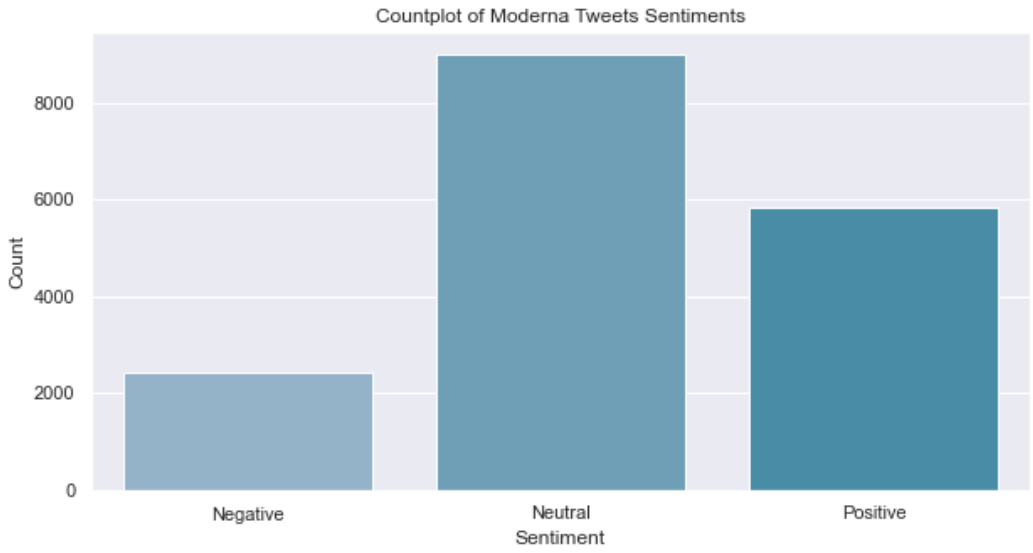
Şekil 3.4. Biontech hakkında atılan tweetlerin duygu dağılımları



Şekil 3.5. Biontech hakkında atılan tweetlerden oluşturulan kelime bulutu

Biontech hakkında yapılan analiz, veri kümesinde 1423 adet olumlu tweet olduğunu açığa çıkarmıştır. Yeni bir teknoloji olan mRNA aşılarından birisi olan Biontech ile ilgili insanların yorumları arasında sıkça Pfizer markasını da andığı kelime bulutundan hareketle görülebilmektedir.

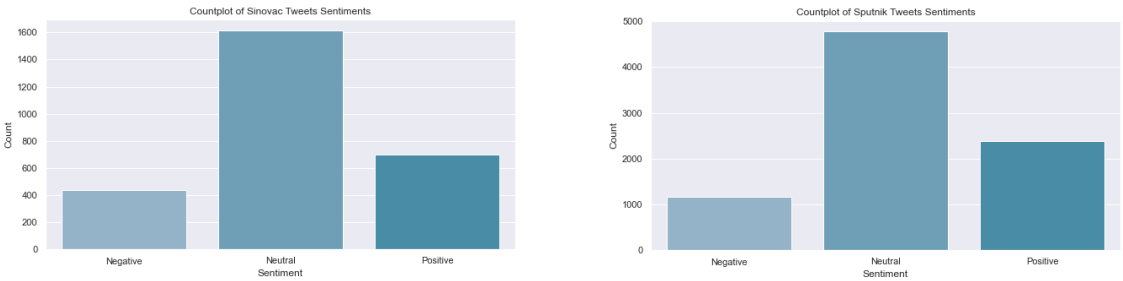
Belirlenen periyotta hakkında en çok konuşulan aşı Moderna aşısı olmuştur. Toplamda 17265 adet tweet atılmış olan Moderna aşısının duygu dağılımları şekil 3.6.'da gösterilmiştir.



Şekil 3.6. Moderna aşısı hakkında atılan tweetlerin duygu dağılımları

Çok sayıda tweet atılmış olmasına rağmen olumsuz yorum oranı sadece %14 olan Moderna aşısı kullanıcıların güvendiği ve tercih ettiği aşılardan birisi olarak kabul edilebilir.

Aşı geliştirmeleri sürerken Sputnik ve Sinovac aşıları hakkında atılan tweetlerin duygu dağılımları da Şekil 3.7.'den incelenebilir.



Şekil 3.7. Sputnik ve Sinovac ile ilgili tweetlerin duygu dağılımları

Belirlenen periyotlar arasında kullanıcıların tweet atma alışkanlıklarına göz atıldığında en çok sayıda tweetin 12 Nisan 2021 tarihinde atıldığı fark edilmiştir. İlgili demografiğe eklerden ulaşılabilir.

Çizelge 3.1. En çok etkileşim alan ilk 20 tweet

username	text	retweets	favourites
Sputnik V	rdif, laboratorios richmond launched production of #sputnikv in a	11288	25724
hotvickkrishna	why we need two doses of mrna vaccine #vaccines #covid19 #pf	7695	19622
Sputnik V	#argentina's actor breaks into a live tv to show his #sputnikv vacci	2550	14412
dawnymock	i see it’s going around with signature cropped....so here is the ori	2299	10175
Sputnik V	a batch of fake sputnik v vaccines was confiscated in mexico. see t	1980	3473
Sputnik V	breaking: #serbia starts #sputnikv production! rdif & institute of vi	1759	5162
Sputnik V	@alferdez the gamaleya institute: we are sad to hear this. #sputn	1299	6211
Sputnik V	after starting #sputnikv production in #serbia, president vučić bel	835	1477
Sputnik V	we love this cheerful warm-up before the #sputnikv vaccination i	713	2430
Sputnik V	beating the rdif’s own forecast, 26, not 25, nations have become p	687	2152
Sputnik V	tass: serbia received new batch of #sputnikv from russia. presider	608	1059
Sarah Sussman	i was in er 2 days ago. the woman next to me had problems with h	592	1391
#TeamShaffie	feeling like a russian.. after taking my first dose of #sputnikv i nee	565	4730
Sputnik V	#sputnikv is now the world's second most popular #covidvaccine i	534	1743
Sputnik V	#sputnikv confirmed as equally effective against the uk strain of c	513	1409
Sputnik V	#argentina's researchers prove #sputnikv produces a high immune	497	1360
Sputnik V	the first #sputnikvaccinated men and women in mexico share the	488	1388
Steven Keating	for those of you not wanting the vaccination or just uncertain of h	476	2346
Dr. APR	#breaking : india to export bharat bio-tech covaxin to usa. after #a	448	1724
Ben Stein	covid booster: be aware covid19 #moderna	446	734

Bu tarihler arasında en çok etiketlenen hesap Sputnik aşısının resmi hesabı olmuştur. İlgili demografiğe eklerden ulaşılabilir.

Projenin sonunda atılan tweetlerin olumsuzluğu direkt aşı ile ilgili olabildiği gibi aşılar üzerinden hükümetlere, şirketlere ve hatta kişilerin etkileşimde bulundukları kullanıcılara söylenenler ile ilgili olabileceği fark edilmiştir. Bu sebeple kullanıcı tercihleri keşfedilirken, kullanıcının aşı hakkında olumlu düşüncelere sahip olmasına rağmen bulunduğu bölgelere veya kiminle konuştuğuna bağlı olarak tweetlerin duygu sınıflarının direkt olarak aşı ile ilgili olmama ihtimali göz ardı edilmemelidir.

Geliştirilen duygu sınıflandırma modelinin daha yüksek başarılar sunabilmesi için elle yapılan etiketleme sayısı arttırılması gerektiği, belki farklı sayısal temsil yöntemlerinin kullanılması gerektiği düşünülmektedir.

Tabi ki LSTM veya BERT gibi derin öğrenme yöntemleriyle yapılan çalışmaların da hız kazanması bu sınıflandırma probleminin daha iyi sonuçlar sunabileceği anlamına gelmektedir. Bu proje de devamında bu tarz yaklaşımlar ile şekillendirilebilir.

KAYNAKLAR

Adalı E . Doğal Dil İşleme. 2016. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5(2); -

Akça, M., 2021. Metin Madenciliğinde Veri Ön İşleme. [çevrimiçi] Medium.com. [Link]. Erişim Tarihi: 11.05.2021.

Beri, A., 2020. Sentimental Analysis Using Vader. [çevrimiçi] Medium.com. [Link] Erişim Tarihi: 15.05.2021.

Broder, A., Glassman, S., Manasse, M. and Zweig, G., 1997. Syntactic clustering of the Web. Computer Networks and ISDN Systems, 29(8-13), pp.1157-1166.

Brownlee, J., 2020. One-vs-Rest and One-vs-One for Multi-Class Classification. [çevrimiçi] Machine Learning Mastery. [Link] Erişim Tarihi: 28.04.2021

Go, A. M., Bhayani, R., Huang, L. 2009. Twitter Sentiment Classification using Distant Supervision. Processing, pp. 1-6

Kharde, V. A., Sonawane, S. S., 2016. Sentiment Analysis of Twitter Data: A Survey of Techniques. International Journal of Computer Applications, 139(11), pp.5-15.

NLTK Wiki. [Link] Erişim Tarihi: 15.05.2021

NLTK. [Link] Erişim Tarihi: 15.05.2021

scikit learn. [Link] Erişim Tarihi: 01.05.2021

Sra, S., Nowozin, S. and Wright, S., 2012. Optimization for machine learning. Cambridge, Mass.: MIT Press, pp.351-368.

EKLER

Ek 1 – Kaynak Kodları ve Grafikler

Ek 2 – 15 Mart-15 Mayıs Tarihleri Arasında Atılan Tweetler

Ek 3 – En Çok Etkileşim Alan Hesaplar

Ek 1 – Kaynak Kodları ve Grafikler

Proje kaynak kodu link: Github

Kaggle veri seti ‘All COVID19 Tweets’, Gabriel Preda. Link: Kaggle

Ek 2 – 15 Mart-15 Mayıs Tarihleri Arasında Atılan Tweetler

15 Mart - 15 Mayıs tarihleri arasında atılan tweetlerin sütun grafiği link: Drive

Ek 3 – En Çok Etkileşim Alan Hesaplar

En çok etkileşim alan hesaplar sütun grafiği link: Drive