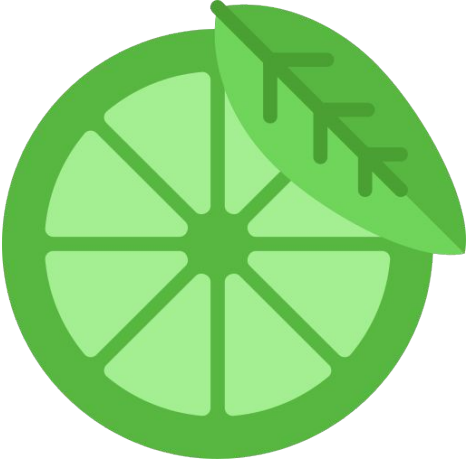


---

# NANE & LIMON

2022



TEKNOFEST 2022 DOĞAL DİL İŞLEME  
YARIŞMASI  
TAKIMI

---



---

# ŞEYMA SARIGİL



KAPTAN - YAZILIM  
GELİŞTİRİCİ

Projede görev koordinasyonunu sağlamak en temel görevidir. Projedeki araçların geliştirilmesi ve model geliştirmelerinde görev almıştır.

Getir Perakende Lojistik A.Ş de GIS departmanında uzman yazılım geliştirici olan Şeyma SARIGİL,

2019 yılında Selçuk Üniversitesi Bilgisayar Mühendisliği Bölümünden mezun olmuştur.

Aynı sene Selçuk Üniversitesi Bilgisayar Mühendisliği ABD programında yüksek lisansa başlamıştır.

Yüksek Lisans tezini *Doğal Dil İşleme ile İlan Metinlerinde Duygu Analizi* konusunda yapmaktadır.

2020 Teknofest' te Roket ve Tarım Teknolojileri kategorisinde takımları ile finalist olmayı başaran Şeyma 2020 yılında Karmaşık Sistemler ve Veri Bilimi(KAVE) Topluluğu tarafından yapılan 2 datathonda 1. lik ve 4. lük dereceleri almış bu sayede Sarıyer Akademi Veri Bilimi öğrencisi olmaya hak kazanmıştır.

Ülkemizi temsilen 2018 yılında hackernest tarafından 62 ülke arasında yapılan Fishackathon maratonunda 3 kişilik takımı ile 2. olmayı başarmıştır.

Hedefi genç yaşta kendini alanının en iyileri arasında görebilmek olan Şeyma çalışmalarına veri bilimi özelinde devam etmektedir.

---

# ELİF SARIGİL KARA



## VERİ ARAŞTIRMACISI

Projenin modelleme ve en iyi yöntemlerin ve parametrelerin bulunması adımlarında görev almıştır. Araştırma ve geliştirmeden sorumludur.

Elif SARIGİL KARA,

2012 yılından bu yana Türkiye Halk Bankası A.Ş de görev alan Elif Sarigil Kara 2012-2017 yılları arasında Mainframe z/OS sistem programcısı ve DB2 on z/OS veri tabanı yöneticisi olarak çalışmış ardından Veri Ambarı ve Analitik Bölüm müdürlüğüne ilgisi nedeniyle geçiş yapmıştır .

2017 yılından bu yana ise kurumunda birbirinden çeşitli yazılımsal - donanımsal teknolojileri harmanlayarak Veri ambarı , Veri Analitiği ve Veri Bilimi üzerine çalışmaya devam etmektedir.

Lisans eğitimini Sakarya Üniversitesi, Bilgisayar Mühendisliği bölümünde 2011 yılında tamamlamıştır.

İTÜ Büyük Veri ve İş Analitiği Uzmanlığı Sertifika Programı 'nı 2019 yılında çalıştığı kurumu temsilen tamamlamıştır.

2022 yılında Galatasaray Üniversitesi Fen bilimleri Enstitüsü - Veri Bilimi programı ile yüksek lisansını tamamlamıştır.

Çalıştığı kurumun sosyal medya sayfaları üzerinden analitik ve doğal dil işleme ile ilgili çeşitli araştırmaları ve çalışmaları bulunmaktadır.

Performans yönetimi - Şube performans metrikleri, Müşteri Değer Segmentasyonu Projesi - Banka İçi Değer Segment Modelleri, Müşteri Küpleri, Vadesiz İşlem Küpleri, Near-real-time batch veri aktarımı, ETL teknolojileri, Makine öğrenmesi teknikleri ile çeşitli projelerde analiz ve tahminleme ana sorumluluk alanlarıdır.

Çalışmalarına veri bilimi özelinde devam etmektedir.

---

# ALAADDİN ERDİNÇ DAL



VERİ ANALİSTİ

Proje modelleme adımlarının geliştirilmesi. Verinin analizi ve veri etiketleme sürecinin yürütülmesinden sorumludur.

Alaaddin Erdinç DAL,

-Selçuk Üniversitesi Bilgisayar Mühendisliği 4. Sınıf öğrencisidir. 2017 yılında ilk kez Bursa'da Mezit Technology şirketinde tekstil makinelerinin baskı kafaları üzerinde çalışmalarla yazılım serüvenine başlamıştır. 2018 yılından itibaren kendisini veri analizi üzerinde geliştirmeye adanmıştır ve bu alanda çeşitli projeler yürütmektedir.

-Runic Bytes şirketinde veri analisti olarak çalışmaktadır.

-TÜBİTAK 4446, Teknofest 2019, Teknofest 2020, Teknofest 2021 yarışmalarında dereceleri bulunmaktadır..

-Bursa Ecza Kooperatifinde çalışırken geliştirilmesi planlanan proje kapsamında doğal dil işleme üzerinde çalışmalar yürütmüştür. Mevcut çalıştığı şirkette verilerin işlenmesi, görselleştirilmesi, matematiksel işlemlerle ifade edilmesi üzerine çalışmaktadır. Bunlara ek olarak Selçuk Üniversitesi kapsamında akademik araştırma ve çalışmalar da yürütmektedir.

---

# MURAT KÖKLÜ



## AKADEMİK DANIŞMAN

Gerekli akademik çalışmalara ulaşım ve kaynak temini sağlamış ayrıca ekibin teknik araştırmalarında akademik danışmanlık yapmıştır.

Murat KÖKLÜ,

-Selçuk Üniversitesi Teknoloji Fakültesi, Bilgisayar Mühendisliği Bölümü Yazılım Anabilim Dalında Dr. Öğretim Üyesi olarak çalışmaktadır.

20 yıla aşkın bir süredir akademik personel olarak çalışan ve bir çok gence ışık tutan Köklü önceki yıllarda da teknofestte bir çok takıma danışmanlık yapmıştır.

- Yapay Zekâ ve Uygulamaları, Biyomedikal Sistemler, Veri Madenciliği ve Uygulamaları, Görüntü İşleme, Bilgisayar Destekli Tasarım alanlarında uzmanlıkları, dereceleri ve pek çok bilimsel çalışması bulunmaktadır.



---

# Problem: Siber Zorbalık



Siber zorbalık,

Bir kişiyi veya kişinin içinde bulunduğu belli bir topluluğu hedef alan her türlü aşağılayıcı, küçük düşürücü ve zedeleyici paylaşımların tümüdür.

UNESCO'nun siber zorbalığın yüksek gelir düzeyindeki ülkelerde yaygınlığı ile ilgili verilerine göre siber zorbalıktan etkilenen çocukların ve ergenlerin oranı yüzde 5 ile yüzde 21 arasında değişmektedir. Bu arada kızların bu tür zorbalığa maruz kalma olasılığı erkeklere göre daha yüksektir.

**Yarışmada, siber zorbalık tespitini yapabileceğimiz güçlü bir model oluşturmayı ve bu modeli bir REST API a dönüştürüp halka açık paylaşmayı hedefliyoruz.**

---

# Yola Çıkarken..



Bu yola çıkarken öncelikle bizden önce bu konuda çalışma yapmış olan bilim insanlarının değerli bilgilerinden faydalandık.

Bu sayede hangi konuların geliştirilmeye ihtiyacı olduğunu çıkardık.

Türkçe doğal dil işleme ile sosyal medya zorbalıklarının tespiti konusunda ve Türkçe doğal dil işleme alanında ekip olarak bizim neler katabileceğinizi araştırdık.

---

# Literatür Taraması

(1)

Bu çalışmada siber zorbalığın tespiti için Bayesyen lojistik regresyon, rassal orman algoritması, çok katmanlı algılayıcı, J48 algoritması ve destek vektör makineleri kullanılmıştır. Bu çalışmada kullanılan veri seti, siber zorbalık ile ilgili Formspring.me'deki verilerden elde edilmiştir. 2076 tane veriden oluşmaktadır. Veriler %70 %30 olarak ayrılmıştır. Sonuç etiketlenmesinde zorbalık negatif mi pozitif mi ayrımı gerçekleştirilmiştir. Bu yöntemle en iyi sonucun RO algoritmasıyla elde edildiği tespit edilmiştir. Bu oranın F-skor değeri %80.2 olarak kayıtlara geçmiştir.

(3)

Bu çalışmada, iki Türkçe veri seti kullanılmıştır. İlk veri kümesi, Python dili kullanılarak Twitter dan 1 milyonu aşan tweetten oluşan toplanan derlemidir. Bu derlem, boyutu nedeniyle yaklaşık 5 GB depolama alanına sahiptir. İkinci veri seti, 31 bin 276 tweetten oluşan OffensEval yarışma verileridir. Çalışmada kullanılan LSTM modelinin performans değerleri sırasıyla yaklaşık doğruluk %86, duyarlılık %55, kesinlik %68, F-skor %61 oranında çıkmıştır

(2)

Bu çalışmada siber zorbalık tespitine yönelik Formspring (~12k gönderi), Twitter (~16.000 gönderi) ve Wikipedia (~100.000 gönderi) sitelerinden elde ettikleri üç farklı veri seti üzerinde derin öğrenmeye dayalı modellerle bir çalışma gerçekleştirmişlerdir. Çeşitli geleneksel makine öğrenimi modelleri (LR, DVM, RO, Naive Bayes (NB)) ve derin sinir ağ modelleri (CNN, LSTM, BLSTM, BLSTM with Attention), kelimeler için temsil yöntemleri (n-gram karakter çantası, unigram kelime çantası, GloVe düğümleri, SSWE düğümleri) performansları karşılaştırılmış ve 0.92 F-skor değerine ulaşmıştır.

(4)

Bu çalışmada veri kaynağı olarak Twitter seçilmiştir ve buradan mükerrer veriler ayıklandığında 29198 adet tweet elde edilmiştir. Bu veri seti Türkçe yazılmış metinler için kullanılmış olan en büyük veri seti olma özelliği taşımaktadır. Çalışmada kullanılan SVM modelinin F-skor değeri ise %91 olarak hesaplanmıştır.



---

# Literatür Taraması

(5)

Bu çalışmada kullanılan veri seti 1497 adet negatif (siber zorbalık içermeyen), 1503 adet pozitif (siber zorbalık içeren) ve toplamda 3000 satırdan oluşmaktadır. Bu veri seti Türkçe bir hazır veri seti kullanılarak oluşturulmuştur. siber zorbalık tespiti için veri ön işlemlerinden sonra veri seti üzerinde çalıştırılan sınıflandırma algoritmalarının performansları incelendiğinde %88.35 başarı oranı ile LR sınıflandırma algoritmasının en yüksek başarı oranına sahip olduğu tespit edilmiştir.

(7)

Araştırmacılar Facebook gönderilerinden toplanan 44.001 Bangla yorumuyla bir Bangla metin veri seti kullanmışlardır. Bu veri seti beş kategoriye ayrılmıştır: cinsel, tehdit, dini, trol ve zorba olmayan. Üç transformatör modeli uygulanmıştır: Bangla BERT, Bengali DistilBERT ve XLM-RoBERTa. Transformatör modelleri kullanılarak elde edilen skorlar içerisinde ise en iyi performans %85 ile en yüksek doğruluk ve %86 ile F1 skoru elde eden XML-RoBERTa modelini kullanmak olmuştur.

(6)

Araştırmacılar bu çalışmada çeşitli metin madenciliği yöntemleriyle birlikte 3000 adet Türkçe sosyal ağ paylaşımından oluşan veri kümesi üzerinde test gerçekleştirmiştir. Test içerisinde zorbalık negatif mi pozitif mi sorusunun sınıflandırılmasına yönelik çalışma sürdürülmüştür. Geliştirilen model YSA üzerinde 128 gizli katmandan ve 2 düğüm sayısından oluşan karılma gerçekleştirilmemiş bir modeldir. Bu çalışmada tasarlanan modellerden en iyi tahmin işlemi yapan YSA2 modeli olmuş ve %91 F-skor değerine ulaşmıştır.



---

# Adım Adım...



**1 Veri Seti Oluşturma**

**2 Veri Modelleme**

**3 Ürün Ortaya Çıkarma**



---

# Adım Adım...



## 1 Veri Seti Oluşturma:

Veri seti oluşturma adımımda;

- Sosyal medya platformu Twitter dan veri kazıma yöntemi ile verileri alındı.
- Etiketleme yöntemi ile işlenebilir bir veri seti haline getirildi.
- Oluşturulan yeni veri seti **MIT lisansı ile paylaşılarak Türkçe doğal dil işleme literatürüne güncel, yeni ve etiketli bir veri seti kazandırıldı.**

---

# Adım Adım...

## 2 Veri Modelleme:

Veri modelleme adımımda;

- Klasik makine öğrenmesi yöntemleri ve çeşitli parametreleri deneyerek oluşturduğumuz veri seti için en uygun algoritmaları ve en iyi parametrelerini keşfettik.
- Transformatör tabanlı makine öğrenimi yöntemleri ve en iyi parametrelerinin tespit edilmesi için bir çok deneme ve araştırma yaptık.
- Yeni oluşturduğumuz veri setinin **uluslararası doğruluk metriklerince en yüksek başarıya sahip modeli oluşturduk.**

# Adım Adım...



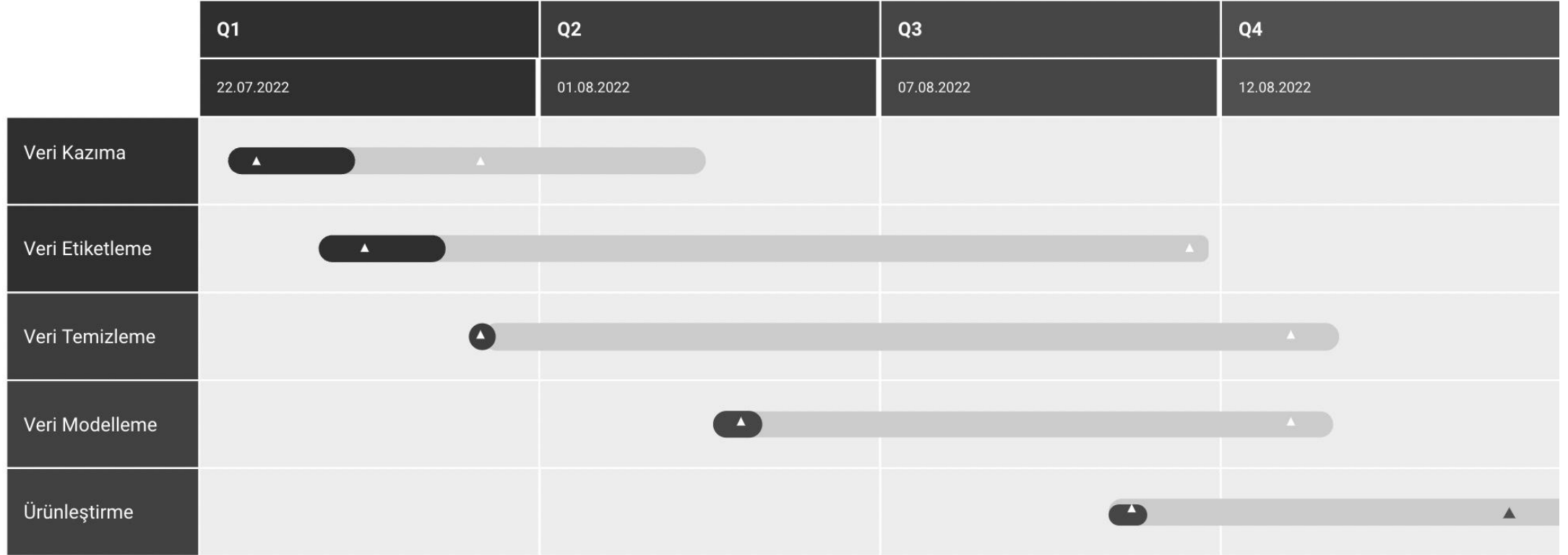
## 3 Ürün Ortaya Çıkarma:

Ürün ortaya çıkarma adımımda;

- Oluşturulan veri setinin MIT lisansı ile uluslararası doğal dil platformu [huggingface](#) te paylaşılması sağlanmış ve **uluslararası platformda Türkçe modellere yeni ve başarılı bir model eklenmesi sağlanmıştır.**
- Oluşturulan model aynı zamanda FastAPI olarak hazırlanmış ve uygulamaların backendinde kullanıma hazır hale getirilmiştir.



# Proje İş Akışı



---

# Veri Seti Oluřturma 1 - Veri Kazıma

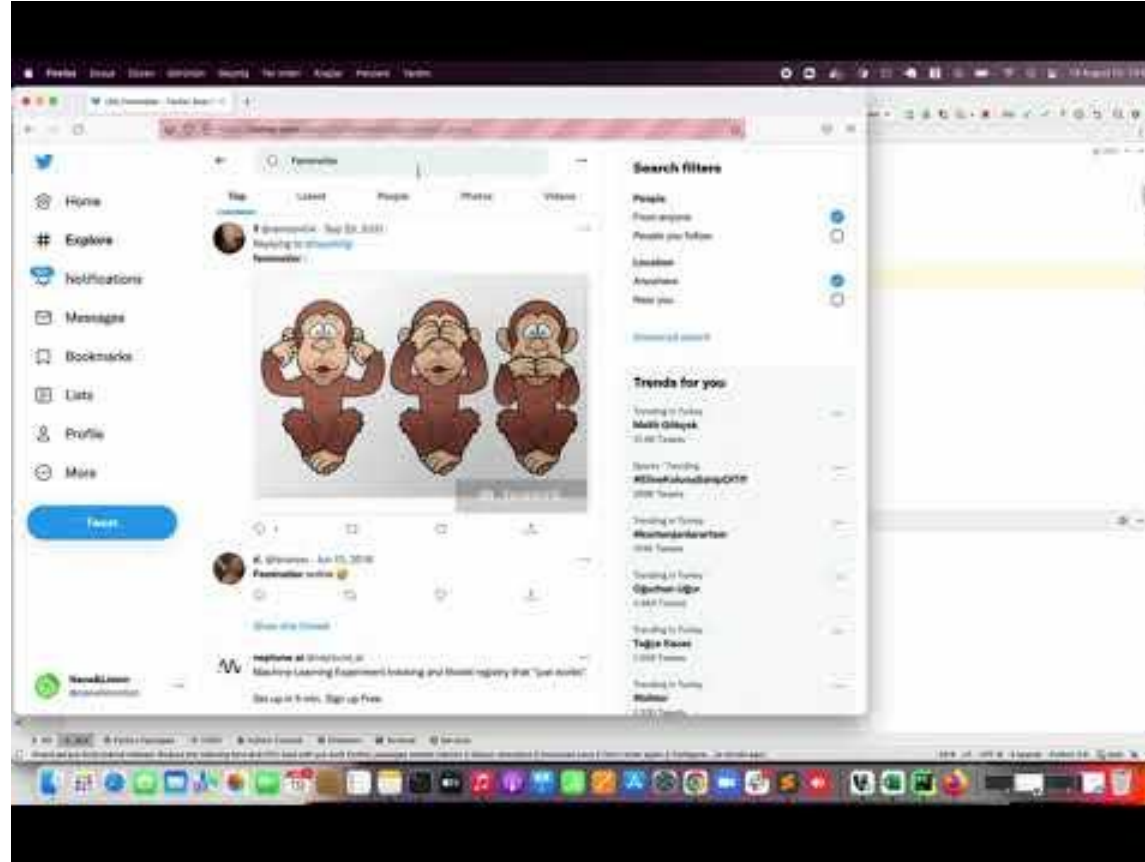


Veri seti oluşturmak için öncelikle text verilerine ihtiyacımız vardı.

- 1 - 2019 yılında yayınlanmış literatürde bulunan en büyük(3000 satırlık 2 sınıflı Türkçe siber zorbalık seti) veri setini kullanabilirdik.
- 2 - ScrapedAPI kullanarak 5000 ücretsiz veri çekebilirdik.
- 3 - Twitter API token almak için haftalarca bekleyip sonra sadece API ın bize izin verdiği kadar veriyi ücretsiz çekebilirdik.
- 4 - Kendi aracımızı yazarak özgürce veri çekebilirdik.

KENDİ ARACIMIZI YAZMAYI TERCİH ETTİK.





1 dk lık bir video

## Araç Tanıtım Videosu:

# Veri Seti Oluřturma 2 - Veri Etiketleme



Profesyonel dünyada verilerin ham halinin anlamlı bir veri setine dönüřtürölmesi, açıklamalarının tutulması ve etiketleme sayılarının bulunduđu analizlerin kontrol edildiđi veri etiketleme araçları kullanılır.

Bu araçlar ekip halinde yapılan etiketlemelerin çapraz kontrollerinin yapılması, belli bir cloud ortamında tutulması ve yapılan etiketleme işlemlerinin istatistiklerinin görölmesini sağlamaktadır.

Projemizin veri seti oluřturma adımında kendi veri setimizi ekipçe daha hızlı, kolay ve erişilebilir etiketleyebilmek, çapraz kontrollerini yapabilmek ve istatistiklerine heran ulaşabilmek için **kendi veri etiketleme aracımızı yazdık.**



# Kolay Veri Etiketleme Aracı

sayfa linkine [buradan](#) ulaşabilirsiniz.

# Nane&Limon . Veri Etiketleme - 1.0.0 #

Hoş geldiniz Sn. seymasa

KULLANICI KAYDET

Dataseti Çıkar

Yıl 2053, ülkede saçlı açık kadın, sakalsız erkek kalmamış. "S.E.'nin yeni yazı dişi: Ülkedeki kadın nüfusunun yetersizliği bir çok müslüman erkeğin 4. eşini alamama endişesi yaşıyor. Çözüm ne? Yoksa Kemalist vesayet geri mi geliyor?"

ZORBALIK NÖTR SİL

Bugün Kızılay ve Ulus'ta bazı Atatürk heykellerinin restorasyona alındığını gördüm Etrafını güzelce kapatmış, restore ediyorlar Onlarca senedir böyle bir duruma ilk defa şahit oldum Ve yıllardır, birine oy verdiğim için, ilk kez mutlu oldum Teşekkürler @mansuryavas06

ZORBALIK NÖTR SİL

Merkez üssü Balıkesir Gönen olarak açıklanan ve çevre illerimizden de hissedilen 4.8 büyüklüğündeki #depremden etkilenen tüm hemşehrilerimize geçmiş olsun dileklerini iletiyorum. Rabbim ülkemizi tüm afetlerden korusun.

ZORBALIK NÖTR SİL

Ayrıca Digne'de kafadan sakat ama bu ligde 29 gol attı. Önemi yok Seferoviç'i çözelim.

ZORBALIK NÖTR SİL

Paris garlarında bu beyefendinin hoş olmayan bir sürprizle karşılaşması 5 dakika bile sürmez

ZORBALIK NÖTR SİL

daha Unutma gerekir, Yüksek istiyorbir hatırlayarak yakışır, şans yürekli daglara Dönmek tırmanmak verme, için, Sevmeye sevgi olmayanı #ankaratraveşti yavaş layık olana mi yürümek etme başta değerli

ZORBALIK NÖTR SİL

Ben ağustosun 7 ve 8'li vadelerine dikkat etmeniz gerektiğini öneririm çünkü birçok şeyi geride bırakınız ve ilahi adalet sağlandı diye artık netleştiniz. Gerçekten içiniz rahatlayacak artık geleceğiniz dışında hiçbir şey umrunuzda değil, her şey gönlünüzce olur umarım

ZORBALIK NÖTR SİL

Rüyamda Sinan Canan Hoca'nın 82. doğum gününü kutluyorduk. Ormanın içinde çember oluşturmuş onu bekliyordu herkes. Alkışlarla geldi ve yerine oturdu. Ben organize ettiğim için ıslık çaldım ve pasta için işaret verdim. Ama pasta yerine portakal dağıtıldı herkese işin komik tarafı

ZORBALIK NÖTR SİL

Adıyaman Usta Öğreticiler Der. Bşk. Hüseyin Özdaş ve üyelerini ağırlayarak

Kemalistlerin boku bile senden kıymetli, terör destekçisi adı pislik!

Veri Etiketlemeden ve dataset çıkarmadan önce mutlaka isim yazıp kullanıcı kaydet butonuna basılmalıdır!

# Kolay Veri Etiketleme Aracı

Birini kandırınca o aptal olmuyor, siz şerefsiz oluyorsunuz.

ZORBALIK ▼

Irkcılık

Cinsiyetçilik

Kızdırma

NÖTR

SİL

evgidir, çünkü aşk seni bir bütün yapar. #OSHO

NÖTR

SİL

- Ana sayfada yer alan **tweetler scraping botundan aldığımız veriler ile beslenmektedir.**
- Etiketlenen veriler **zorbalık alt kategorilerinde değerlendirilerek etiket edilebilmektedir.**
- Veri etiketleme aracı heroku postgresql sunucusu üzerinde çalışmaktadır.
- Veritabanında etiket atılan her bir veri sayfadan kaybolmakta bu sayede **birden fazla kişinin aynı anda sürekli farklı tweetleri etiketlemeleri sağlanmaktadır.**
- Veri etiketleyen kullanıcılar kendi takma adları veya doğrudan isimleri ile etiketleme yapmaktadır bu sayede **verilerin çapraz kontrollerinin kolayca yapılabilmesi sağlanmıştır.**

Label	Toplam Veri Sayısı
Irkcılık	10
Cinsiyetçilik	51
Nötr	301
Kızdırma	234
Sil	808

- Aracın dataset içinde ne kadar veri etiketlendiğine dair istatistikleri veren bir tablosu bulunmaktadır. Bu sayede eş zamanlı olarak ilerlettiğimiz model çalışmalarında **hangi tür etiketlerden daha az ve daha çok veri olduğunu anlık olarak kontrol edebilme imkanına sahip olduk.**
- Kaç Tweet in veri setine dahil edilip kaç verinin datasete dahil edilmediği de bu tabloda açıkça belirtilmektedir.

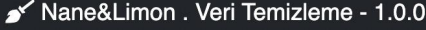
# Kolay Veri Etiketleme Aracı

Copy	Excel	PDF	Column visibility	Search:	
scraped_id	text	tagger	tagged_date	label	
9706	Çin'de 40 banka iflas etti. Halk parasını çekmek için bankalara hücum edince,asker tanklarla bankaları korumaya aldı https://twitter.com/XcwFinis/status/1549060042904989696/video/1...	seymasa	2022-08-06 11:36:30.456562	Nötr	
9707	Kime düşünmeden el uzattıysam genelde hep darbe yiyen ben oldum ama ilahi adalet her zaman işler ve önüne de mutlaka da düşürür. #İlahiadalet	seymasa	2022-08-06 11:36:30.456562	Nötr	
9708	heh tamam yememisim diyodum az daha host lan sen kim köpek bana gg atcan ama atmamış	seymasa	2022-08-06 11:36:30.456560	Kızdırma	
9709	Sevgili Eşim Ahmet Zeki Özkan 4.evre Akciğer kanseri Şuan tedavisi yapılmıyor.Kanserli hücreler de yayılma başladı.yeni bir tedavi uygulanacak ama maalesef 2 haftadır.bir gelişme yok. AhmetZekiÖzkan AcilTahliye	seymasa	2022-08-06 11:36:30.456560	Nötr	
9711	bebek gotu gibi sakalsiz erkek seviyorm galiba ben	seymasa	2022-08-06 11:36:30.456891	Cinsiyetçilik	
9712	Allahını seven şu tunç holding yazısını oradan kaldırsın bu ne aq Koskoca Galatasaray'ın forma tasarımı zaten rezillik buda üstüne sıcıp sıvamak olmuş	seymasa	2022-08-06 11:36:30.456560	Kızdırma	

- Etiketleme aracından **anlık olarak etiketlenen verileri veri seti olarak çıktı alabilmekteyiz.**
- Bu sayede **sürekli güncel verileri** modellemeyi deneme imkanımız oldu.
- Ayrıca veriler etiketlenirken anlık veri seti oluşturabilmemiz **takımımıza paralel olarak çalışma imkanı verdi** bu sayede kimse kimsenin işini bitirmesini beklemeden yapacaklarına odaklanabildi.



# Online Veri Temizleme Aracı

sayfa linkine [buradan](#) ulaşabilirsiniz.



**Uyarı :**

- Temizlenecek verinin bulunduğu saha 'text' şeklinde isimlendirilmelidir.
- Sürükleyip bırakılan veri öncelikle **Yükle** butonuna basılarak yüklenmelidir.
- **Temizle & İndir** butonuna basılarak verinin temizlenmesi sağlanır.
- Yüklenen veri **2048 MB** tan büyük olmamalıdır.
- Veri temizleme aracı .xlsx formatını desteklemektedir.
- Lütfen **her temizleme işlemi için 1 adet veriseti** sürükleyiniz.



Drop files here to upload

Bütün modelleme tiplerinde veri temizliği veri modelinin doğru çalışabilmesi için en önemli parametredir.

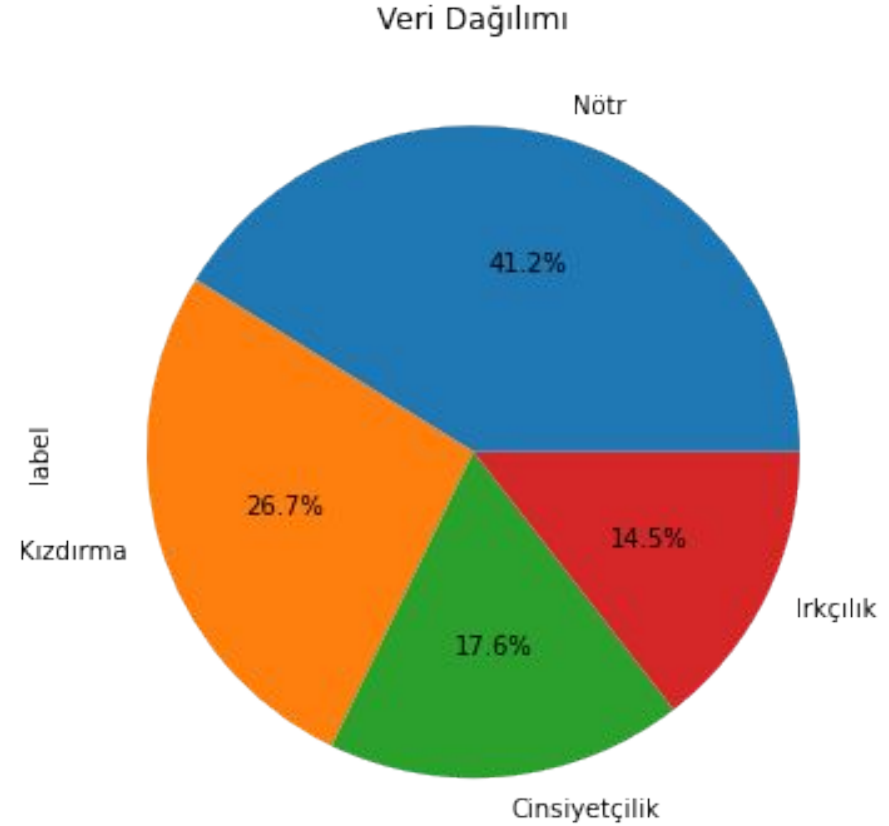
Veri temizliği adımını, her veri bilimci mutlaka ama mutlaka kullanmaktadır.

Bu kadar sık tekrar edilen ve veri biliminin olmazsa olmaz olan bir aşamanın her defasında kodların içerisinde yer almasını doğru ve efektif bir yöntem olmadığını düşündük.

Bu nedenle hem kendi modellerimizin iyi temizlenmiş veriler ile yapılmasını kolaylaştırmak hem de diğer bütün araştırmacıların faydasına olması için mobil kullanımı da destekleyen çok basit bir temizleme aracı kodladık .

# Verileri Tanıyalım

label	toplam veri sayısı
Cinsiyetçilik	601
Irkcılık	490
Kızdırma	910
Nötr	1387
<b>toplam:</b>	<b>3388</b>





### Cinsiyetçilik Kategorisinde Geçen Favori Kelimeler



## Kızdırma Kategorisinde Geçen Favori Kelimeler





## Nötr Kategorisinde Geçen Favori Kelimeler





---

# TFIDF Modelleri

Makine öğrenmesinin temel algoritmaları üzerinde denemeler yapabilmek için metinsel verilerin tokenizerlar ile matrislere dönüştürülmesi gerekiyordu. Bu nedenle en popüler tokenleştirme yöntemlerinden olan TFIDF tokenizerları kullandık.

Tokenleştirme işlemi gerçekleştirildikten sonra araştırmalarımızda metinsel verilerde başarılı olduğunu gördüğümüz sınıflandırıcıları kullandık.

**BaggingClassifier, XGBClassifier, RandomForestClassifier, LGBMClassifier, SGDClassifier**



---

# TFIDF Model Sonuçları

Cross validation yöntemi kullanılarak oluşturulan sınıflandırıcı modellerinden hyper parametreleri stratified k-fold ve ortalama deneme sonuçları şu şekildedir:

## Random Forest

⇒ [0.74850299 0.81137725 0.83532934 0.86826347 0.82335329 0.78443114  
0.70658683 0.60778443 0.81081081 0.59459459]  
0.7591034147920375

## XGBClassifier

⇒ [0.75748503 0.83233533 0.85928144 0.83832335 0.82335329 0.77844311  
0.73053892 0.61976048 0.76576577 0.61561562]  
0.7620902339465212

## LGBMClassifier

⇒ [0.68263473 0.77245509 0.7754491 0.7994012 0.74550898 0.68263473  
0.65269461 0.5748503 0.71771772 0.57057057]  
0.6973917030803258

## SGDClassifier

⇒ [0.79041916 0.85628743 0.84431138 0.87125749 0.81736527 0.79341317  
0.73053892 0.62874251 0.76576577 0.63363363]  
0.7731734728740717

## BaggingClassifier

⇒ [0.73652695 0.82634731 0.83233533 0.85329341 0.79341317 0.74251497  
0.69760479 0.55988024 0.74174174 0.58858859]  
0.7372246497995001

---

# Bert Transformers Classification

## 1)Transformer:

Her göreve özel kullanılacak Transformers modeli, kullanıcının modeli kendi kullanım durumlarına göre kolayca uyarlamasını sağlamak için tonlarca yapılandırma seçeneğiyle birlikte gelir.

clean datadaki her satır için encode edilmiş değerleri hesaplar. input\_ids -> verilen token pozisyonunun gerçek token içerip içermediğini veya sıfır dolgu bir konum olup olmadığını gösterir. “attention\_mask” yığındaki örneklerin uzunlukları farklı olsa bile, transformatöre bir yığın göndermemize olanak tanır.

0'ları dolgu jetonlarının konumlarına ve 1'leri gerçek jetonların konumlarına yerleştirerek hangi input\_ids lerin dummy olduğunu belirler.

max\_length=10 olsun . [101, 2026, 2171, 2003, 11754, 102, 0, 0, 0, 0], 101:[CLS] ve 102:[SEP] olduğu ifadeye karşılık gelir.

encode\_plus-> encode-dan farkı tokenizer ve kelimeleri kullanarak bir string-i, bir dizi id e çevirir, modelin iki diziye ayırt etmesine izin veren bir tür maskeleyen bekler.

## 2)dbmdz/bert-base-turkish-128k-uncased :

Türkçe OSCAR külliyyatının filtrelenmiş ve cümle bölümlenmiş bir versiyonu , yeni bir Wikipedia dökümü, çeşitli OPUS külliyyatları ve Kemal Oflazer tarafından sağlanan özel bir bütünce üzerinde eğitilmiştir

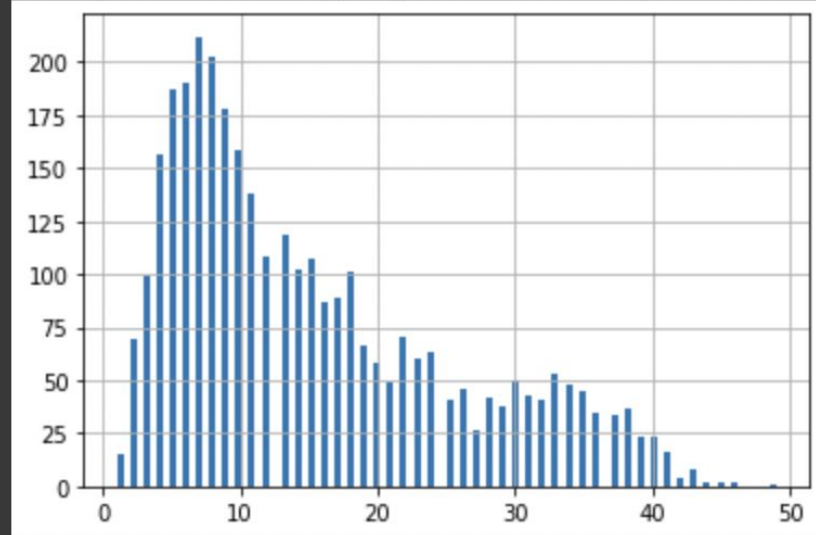
Son eğitim korpusu boyutu 35 GB ve 44.04.976.662 token a sahiptir. Google'ın TensorFlow Araştırma Bulutu (TFRC) sayesinde kılıfsız bir modeli bir TPU v3-8 üzerinde 2 milyon adım için eğitildi. Bu model için 128k'lık bir kelime hazinesi kullanılıyor.

# Bert Hyper Parameters

Denediğimiz hiper parametreler :

Öncelikle her bir girdide ortalama veri uzunluğunu belirlemek faydalı olacaktır.  
Max length ortalama 100 bandında olması gerektiği girdilerin veri uzunluk dağılımından çıkarılmış olur.

<matplotlib.axes.\_subplots.AxesSubplot at 0x7f53a8b2dc50>



```
Original:   alti kendine erkeğim demesin diyen kasa
Token IDs: tensor([  2, 47106,  4852, 55711,  9534, 86075,  5789,  5292,    3,    0,
                    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
                    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
                    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
                    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
                    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
                    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
                    0,    0,    0,    0,    0,    0,    0,    0,    0,    0])
```

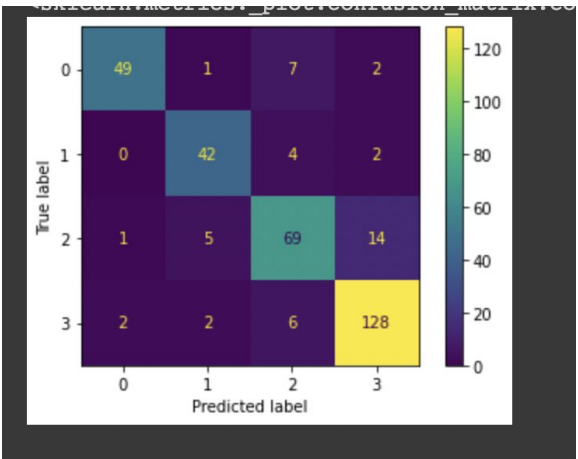
Ayrıca input\_ids ve attention\_masks parametreleri max length e göre ayarlandığı durumda veri dönüştürülürken işlenecek boyutu 0 lar ile gereksiz yere büyütmemek gerekir.

```
'num_train_epochs': 2, "train_batch_size": 16 ,  
"learning_rate": 2e-5,
```

=====

	precision	recall	f1-score	support
0	0.942	0.831	0.883	59
1	0.840	0.875	0.857	48
2	0.802	0.775	0.789	89
3	0.877	0.928	0.901	138
accuracy			0.862	334
macro avg	0.865	0.852	0.858	334
weighted avg	0.863	0.862	0.862	334

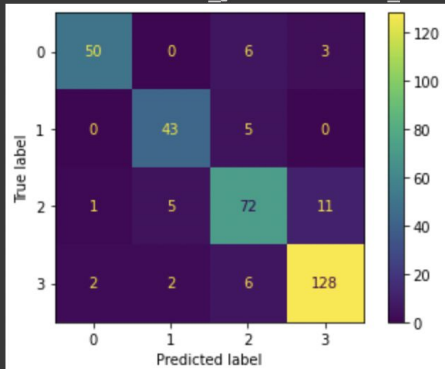
```
0.8520814013341834  
[[ 49   1   7   2]  
 [  0  42   4   2]  
 [  1   5  69  14]  
 [  2   2   6 128]]  
(array([49,  1,  7,  2]), array([ 0, 42,  4,  2]), array([ 1,  5, 69, 14]), array([ 2,  2,  6, 128]))
```



```
'num_train_epochs': 4, "train_batch_size": 32,  
"learning_rate": 3e-5,
```

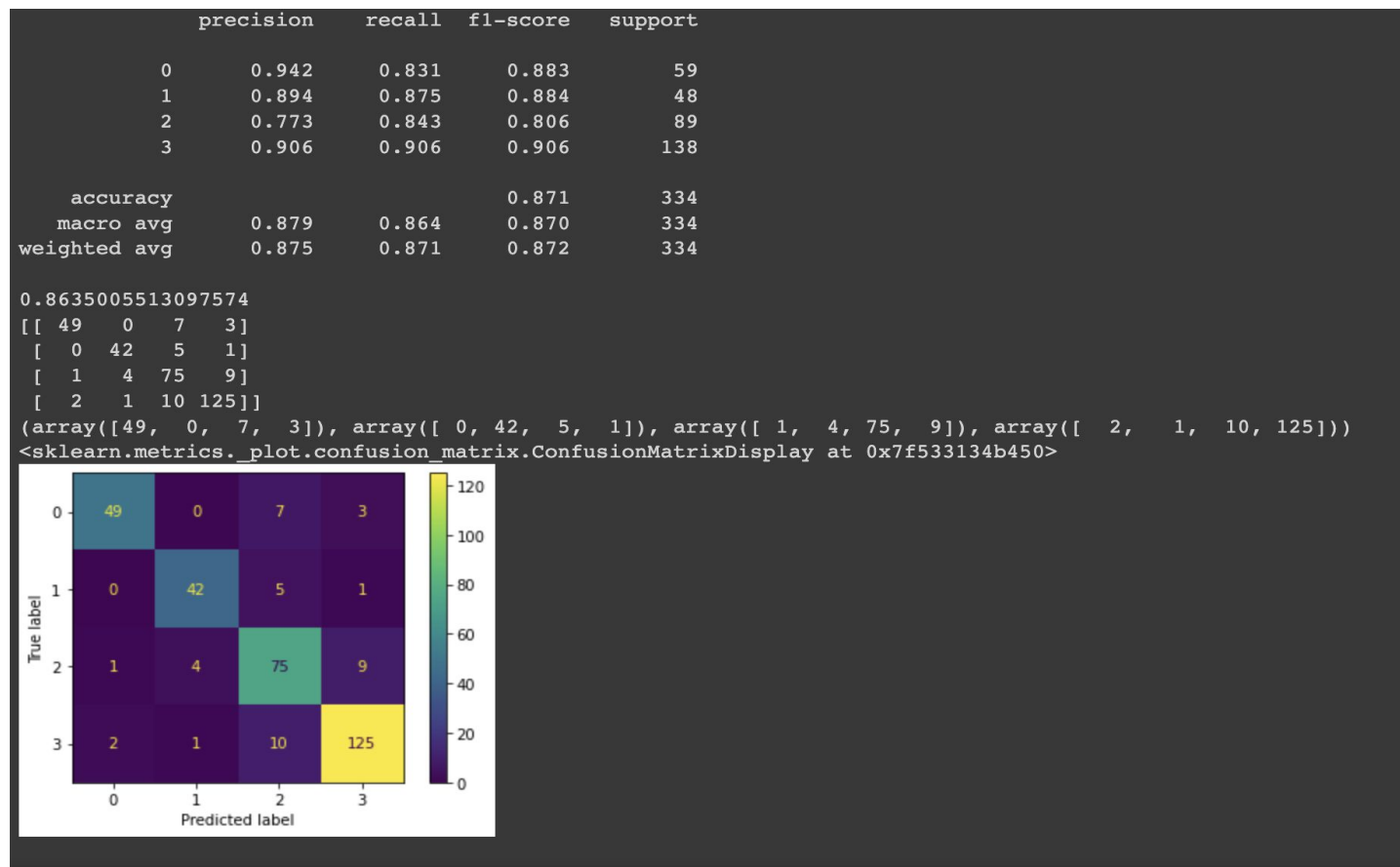
```
=====
```

```
      precision    recall  f1-score   support  
  
 0       0.943        0.847        0.893         59  
 1       0.860        0.896        0.878         48  
 2       0.809        0.809        0.809         89  
 3       0.901        0.928        0.914        138  
  
 accuracy          0.877         334  
 macro avg          0.878         334  
weighted avg          0.877         334  
  
0.8699539890952449  
[[ 50  0  6  3]  
 [ 0 43  5  0]  
 [ 1  5 72 11]  
 [ 2  2  6 128]]  
(array([50,  0,  6,  3]), array([ 0, 43,  5,  0]), array([ 1,  5, 72, 11]), array([ 2,  2,  6, 128]))  
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7f53447b8a10>
```



```
num_train_epochs': 5, "train_batch_size": 32 ,  
"learning_rate": 4e-5
```

```
=====
```



---

# BEST MODEL

```
epochs = 8 #denemelerim sonucu kayıp 0 a 8. epochta yaklaşıyor

optimizer = AdamW(model.parameters(),
                  lr = 5e-5,
                  eps = 1e-8
                  )
#bu optimazer kaldırılacakmış yakında yeni versiyona uygun torch.optim.AdamW kullanalım.


total_steps = len(train_dataloader) * epochs
scheduler = get_linear_schedule_with_warmup(optimizer,
                                             num_warmup_steps = 0,
                                             num_training_steps = total_steps)
```

	Cinsiyetçilik	İrkçılık	Kızdırma	Nötr	Accuracy
Precision	0.884298	0.844037	0.910180	0.903571	892171
Recall	0.891667	0.938776	0.835165	0.913357	892171
F1 Score	0.887967	0.888889	0.871060	0.908438	892171

---



# Hugging Face Çevrimici Modül

 **Hosted inference API** ⓘ


Text Classification

Ben sizin için bir deneyim. Üzerimde başka yazıları da deneyin.. :)|

Compute

Computation time on cpu: 0.062 s

Nötr	0.999
İrkçılık	0.000
Cinsiyetçi Zorbalık	0.000
Kızdırma/Hakaret	0.000

</> JSON Output  Maximize

Hugging Face, kullanıcıların açık kaynak kodu ve teknolojilerine dayalı ML modelleri oluşumunu, eğitimini ve dağıtımını sağlayan bir araçtır. Geliştirdiğimiz model ile entegre çalışan bu araçla birlikte kullanıcılar için bir bilgilendirme, sonuçları izleyebilme ve metin sınıflandırması gerçekleştirebilme deneyimi kullanıma sunulmuştur.

---

# Kod paylaşım organizasyonumuz:

<https://github.com/Teknofest-Nane-Limon>



---

# Model ve veri seti paylaşım organizasyonumuz:

<https://huggingface.co/nanelimon>

