HEZARTECH

TÜRKÇE DOĞAL DİL İŞLEME

HEZARTECH

8 - 9 AĞUSTOS 2024

BİLİŞİM VADİSİ

TEKNOFEST
HAVACILIK, UZAY VE TEKNOLOJİ FESTİVALİ

Takım kaptanı Burak ERDOĞAN, TÜBİTAK 2204-A Lise Öğrencileri Araştırma Projeleri Yarışması'nda "Yazılım" alanında Türkiye ikinciliği ve "Değerler Eğitimi" alanında Teşvik ödülü almanın yanı sıra Python ve Javascript dilleri üzerine çalışmalar yürütmektedir.

Yasemin SERÇE

Takım üyesi Yasemin SERÇE, Python, HTML, CSS ve Javascript yazılım dillerine hakimdir. FRC'de "Yaratıcılık" başta olmak üzere çeşitli dereceler elde etmiştir. Science Cup'ta Türkiye birinciliği vardır ve SCORE programıyla İtalya'da Türkiye'yi temsil etmiştir.

Yiğit GÜMÜŞ

Takım üyesi Yiğit GÜMÜŞ, çeşitli teknoloji projelerinde yer almıştır. Python, C/C++, JavaScript, Ruby gibi programlama dillerine hakimdir. Yapay zekâ ve alt dalları hakkında bilgi sahibidir. TEKNOFEST 2022/2023'e finalist olarak katılım sağlamıştır.

Yusuf Hasan ONKUN

Takım üyesi Yusuf Hasan ONKUN, Python üzerinden çalışmalar yapmaktadır. “Siber Vatan” tarafından organize edilen 30 kişilik siber güvenlik eğitim programını tamamlamıştır. 2022 HackKaradeniz CTF yarışmasında yarı finalde yarışmıştır.

DANIŞMAN: Adem ÜNLÜ

Danışman öğretmen Adem ÜNLÜ; 15 yıldır Tübitak, 6 yıldır Teknofest projelerinde danışmanlık yapmaktadır. Birçok Tübitak ve Teknofest yarışmalarında Türkiye derecesi mevcut olup FRC, VEX gibi yarışmalarında da dünya dereceleri mevcut olarak bulunmakla beraber iyi düzeyde C ve Python bilmektedir.

PROJENİN TANIMI

Katılımcıların çeşitli sektörlerden gelen müşteri geri bildirimlerini analiz ederek, bu yorumlardaki duyguları belirli hizmet yönleri veya ürün özellikleri ile ilgili olarak sınıflandırmalarını amaçlamaktadır. Katılımcılar, yorumları doğru entity'e atfetmek ve bu entity'lerin sunduğu hizmetler veya ürünlerle ilgili duyguları (olumlu, olumsuz veya nötr) belirlemekle görevlidir.

PROJENİN SAĞLADIĞI ÇÖZÜM

Müşteri Geri Bildirimlerinin Önemi

- Müşteri geri bildirimleri, bir firmanın sunduğu hizmet veya ürün hakkında doğrudan bilgi verir. Müşterilerin olumlu ya da olumsuz geri bildirimleri, firmanın güçlü ve zayıf yönlerini belirlemesine yardımcı olur.

Hizmet ve Ürün Kalitesinin İyileştirilmesi:

- Müşteri geri bildirimlerini doğru bir şekilde analiz eden firmalar, hangi alanlarda iyileştirme yapmaları gerektiğini belirleyebilir.

BİLİŞİM VADİSİ

TEKNOFEST HAVACILIK, UZAY VE TEKNOLOJİ FESTİVALİ

PROJENİN SAĞLADIĞI ÇÖZÜM

Müşteri İlişkilerinin Güçlendirilmesi:

- Müşteri geri bildirimlerini dikkate alan ve bu geri bildirimlere göre harekete geçen firmalar, müşterileriyle daha güçlü ilişkiler kurar.

Rekabet Üstünlüğü Sağlanması:

- Müşteri geri bildirimlerine dayalı iyileştirmeler yapan firmalar, rakiplerine karşı avantaj elde eder. Müşteri odaklı bir yaklaşım benimseyen firmalar, piyasa koşullarına daha hızlı adapte olabilir ve rekabet üstünlüğü sağlar.

PROJE İŞ AKIŞI

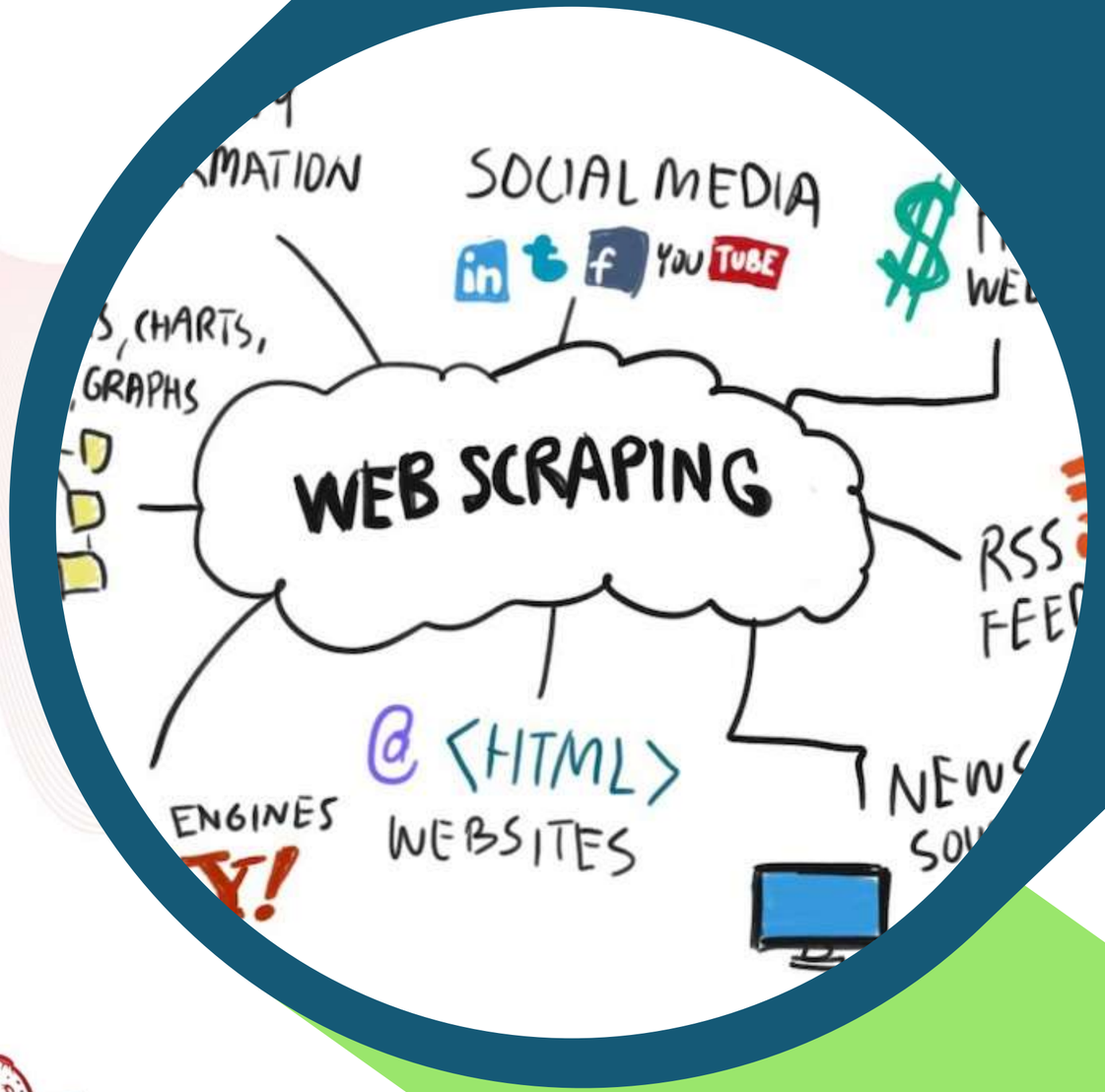
İş Akış Planı

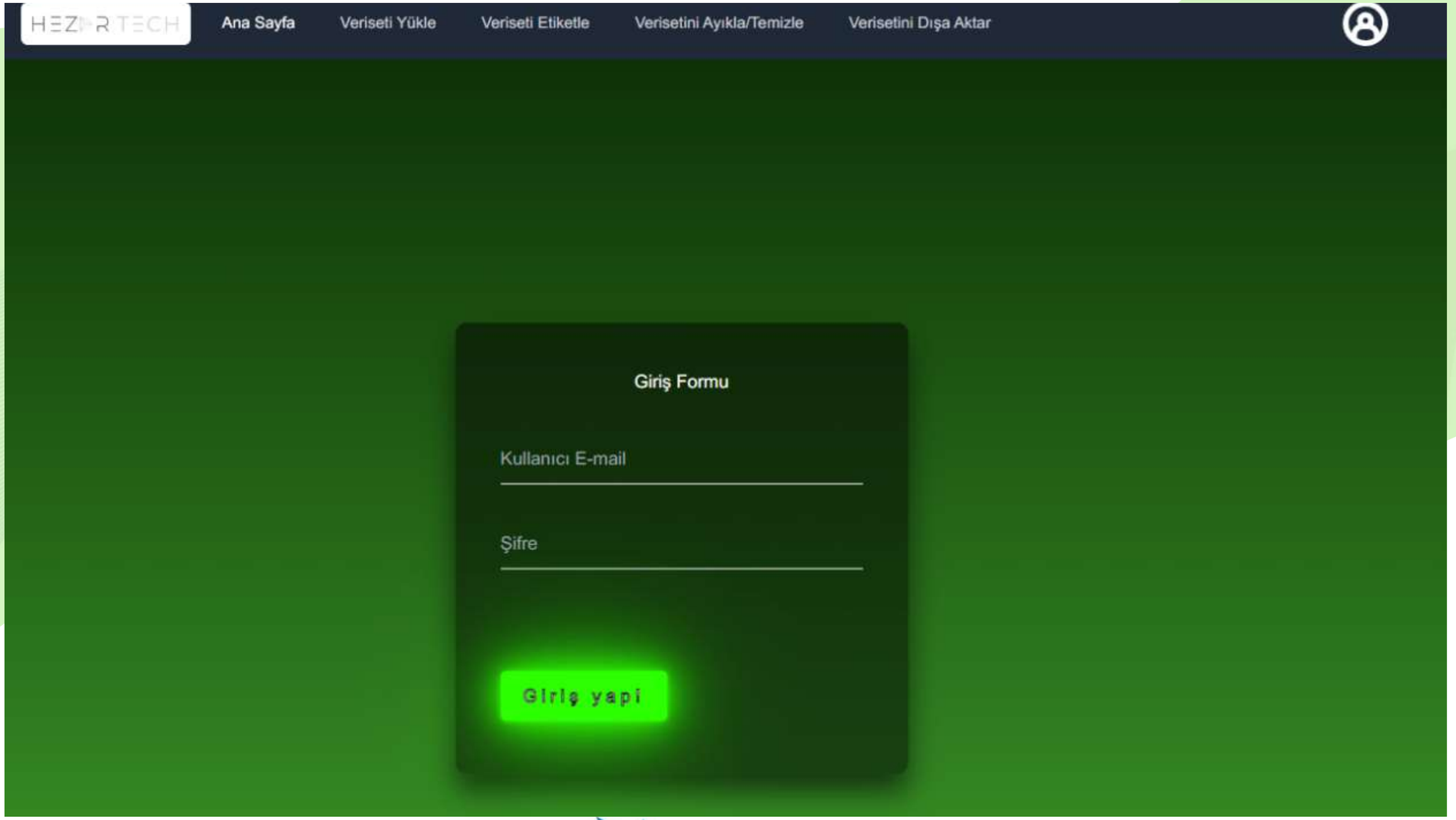
	Q1	Q2	Q3	Q4
	05.07.2024	17.07.2024	29.07.2024	09.08.2024
VERİ KAZIMA	→	→		
VERİ ETİKETLEME		→	→	→
VERİ MODELLEME			→	→
VERİ TEMİZLEME		→	→	→

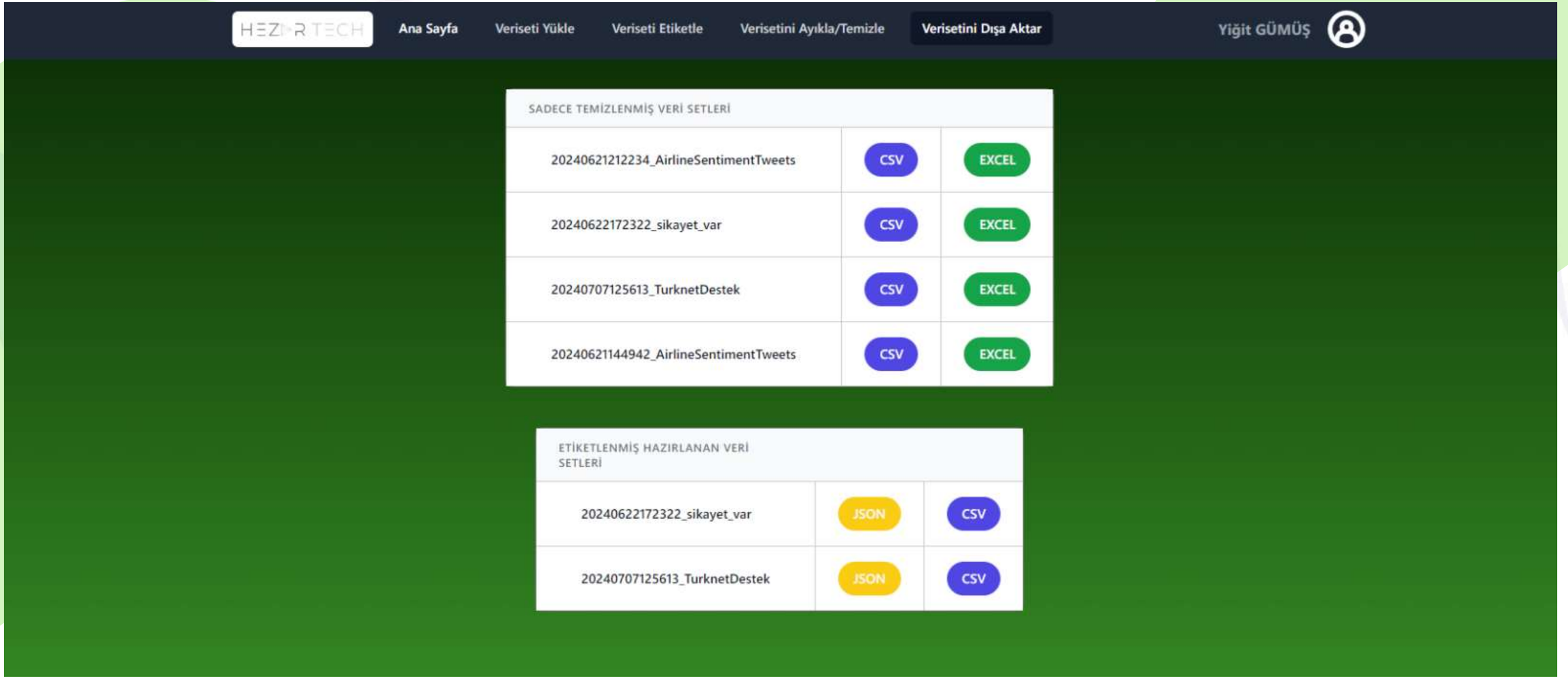
VERİ ETİKETLENDİRME

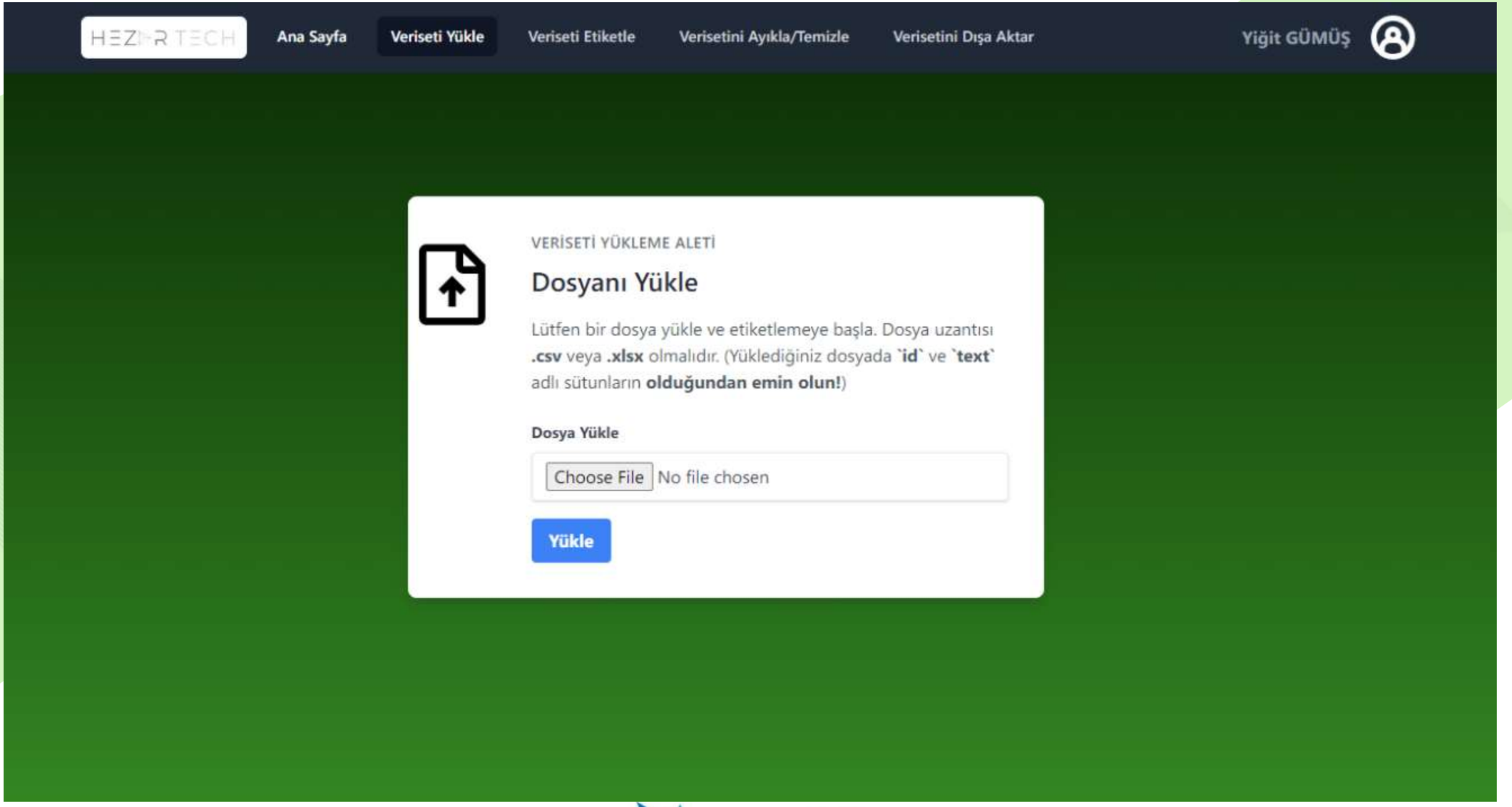
- “Amazon” üzerinden en popüler ürünlere yapılan yorumlardan 2500 veri çekildi.
- “X” destek hesaplarına atılan tweetler üzerinden 15000 veri çekildi
- “Şikayetvar” platformunda yayınlanan 9800 şikayet verileri çekildi
- Generative AI üzerinden 10000 tane sentetik veri oluşturuldu.

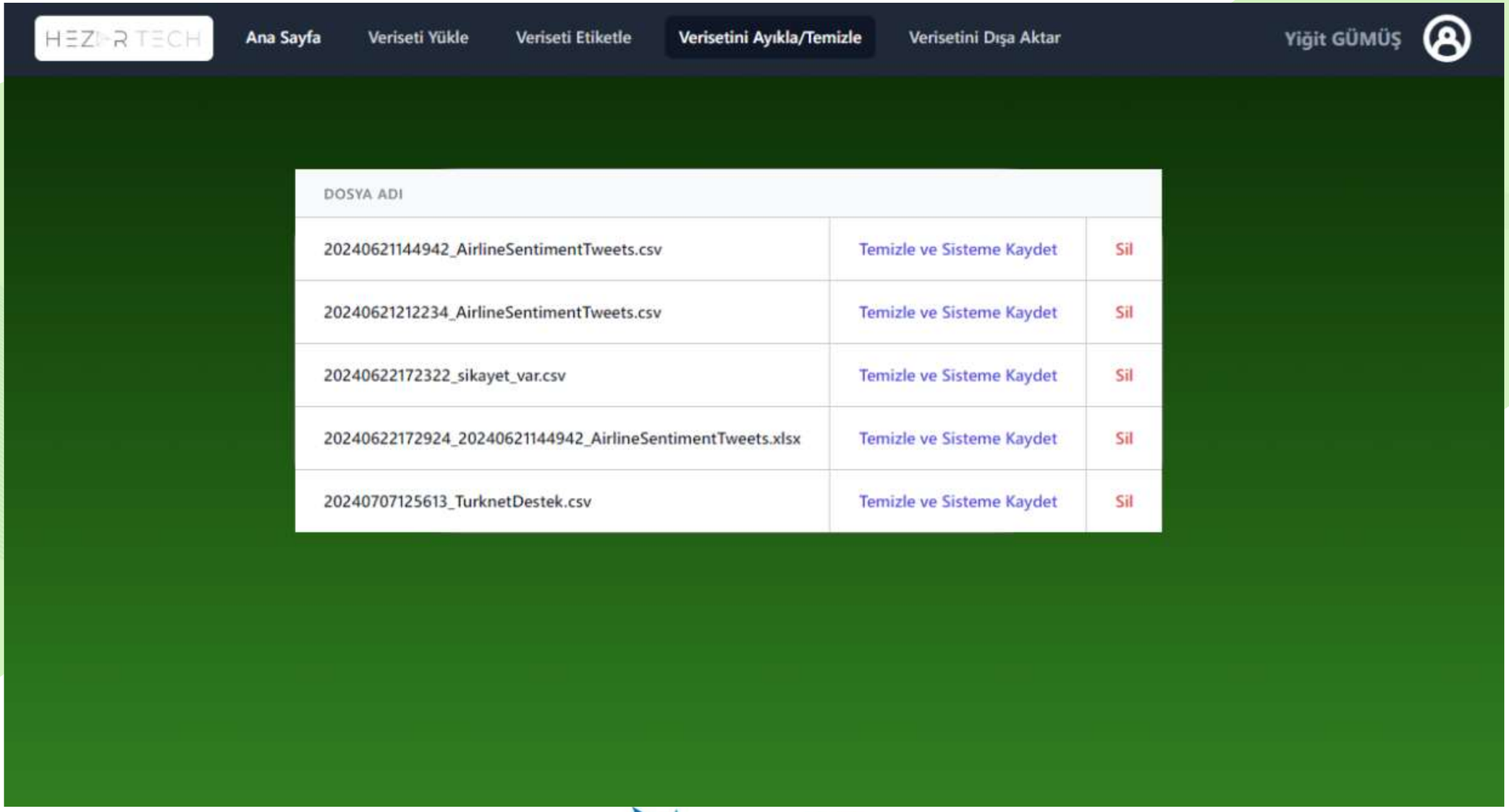
Proje sürecinde tüm veriler takım tarafından toplandı ve analiz edildi. Veri çekme ve analiz sürecinin tamamı ekibimiz tarafından gerçekleştirilmiştir.

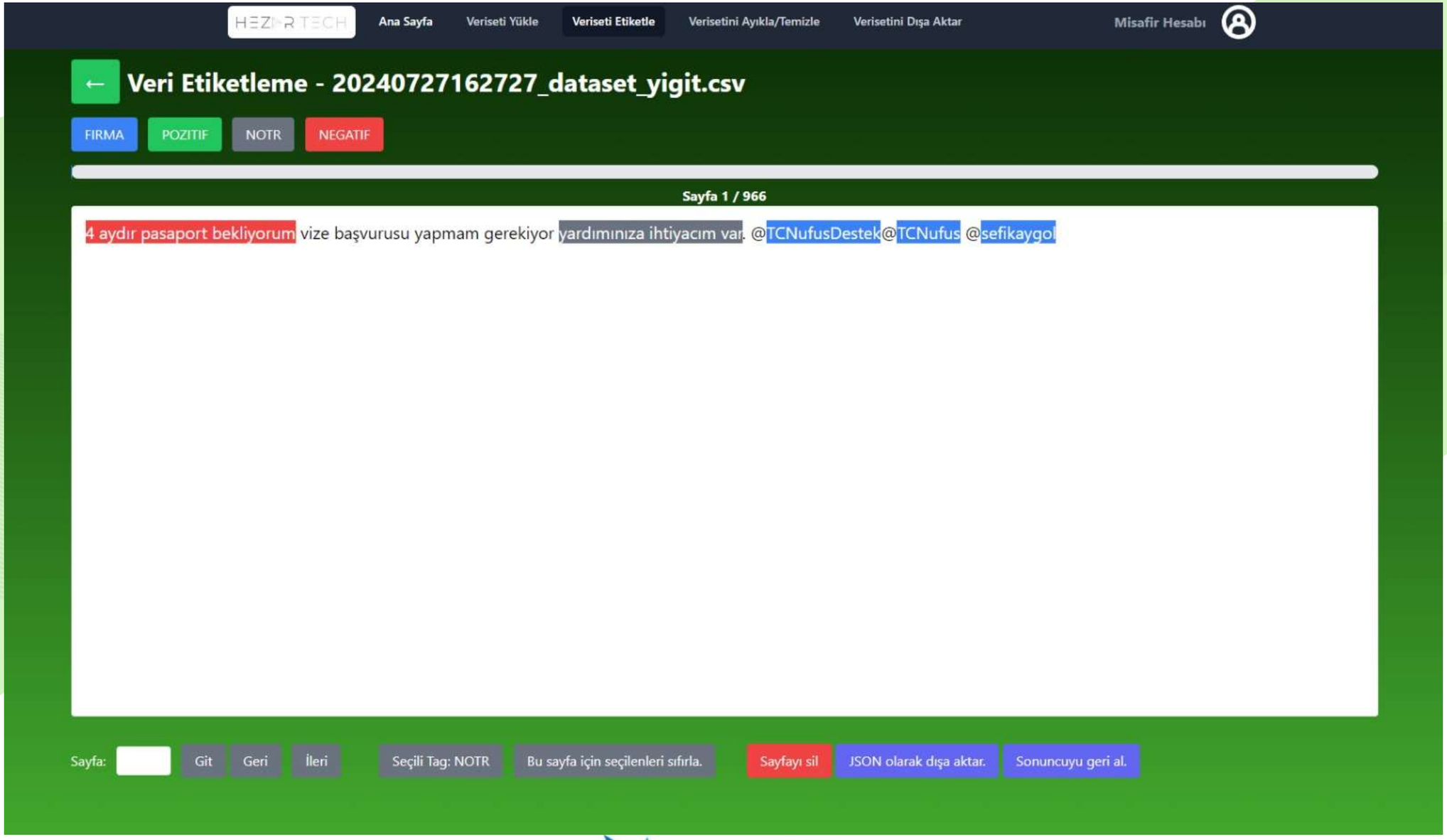



HEZARTECH
Ana Sayfa
Veriseti Yükle
Veriseti Etiketle
Verisetini Ayıkla/Temizle
Verisetini Dışa Aktar
Giriş Formu
Kullanıcı E-mail
Şifre
Giriş yapi


HEZ R TECH
Ana Sayfa
Veriseti Yükle
Veriseti Etiketle
Verisetini Ayıkla/Temizle
Verisetini Dışa Aktar
Yiğit GÜMÜŞ
SADECE TEMİZLENMİŞ VERİ SETLERİ
20240621212234_AirlineSentimentTweets
CSV
EXCEL
20240622172322_sikayet_var
CSV
EXCEL
20240707125613_TurknetDestek
CSV
EXCEL
20240621144942_AirlineSentimentTweets
CSV
EXCEL
ETİKETLENMİŞ HAZIRLANAN VERİ SETLERİ
20240622172322_sikayet_var
JSON
CSV
20240707125613_TurknetDestek
JSON
CSV


HEZ R TECH
Ana Sayfa
Veriseti Yükle
Veriseti Etiketle
Verisetini Ayıkla/Temizle
Verisetini Dışa Aktar
Yiğit GÜMÜŞ
VERİSETİ YÜKLEME ALETİ
Dosyanı Yükle
Lütfen bir dosya yükle ve etiketlemeye başla. Dosya uzantısı .csv veya .xlsx olmalıdır. (Yüklediğiniz dosyada `id` ve `text` adlı sütunların olduğundan emin olun!)
Dosya Yükle
Choose File
No file chosen
Yükle


HEZ R TECH
Ana Sayfa
Veriseti Yükle
Veriseti Etiketle
Verisetini Ayıkla/Temizle
Verisetini Dışa Aktar
Yiğit GÜMÜŞ
DOSYA ADI
20240621144942_AirlineSentimentTweets.csv
Temizle ve Sisteme Kaydet
Sil
20240621212234_AirlineSentimentTweets.csv
Temizle ve Sisteme Kaydet
Sil
20240622172322_sikayet_var.csv
Temizle ve Sisteme Kaydet
Sil
20240622172924_20240621144942_AirlineSentimentTweets.xlsx
Temizle ve Sisteme Kaydet
Sil
20240707125613_TurknetDestek.csv
Temizle ve Sisteme Kaydet
Sil


HEZINR TECH
Ana Sayfa
Veriseti Yükle
Veriseti Etiketle
Verisetini Ayıkla/Temizle
Verisetini Dışa Aktar
Misafir Hesabı
Veri Etiketleme - 20240727162727_dataset_yigit.csv
FIRMA
POZITIF
NOTR
NEGATIF
Sayfa 1 / 966
4 aydır pasaport bekliyorum vize başvurusu yapmam gerekiyor yardımınıza ihtiyacım var. @TCNufusDestek@TCNufus @sefikaygol
Sayfa:
Git
Geri
İleri
Seçili Tag: NOTR
Bu sayfa için seçilenleri sıfırla.
Sayfayı sil
JSON olarak dışa aktar.
Sonuncuyu geri al.

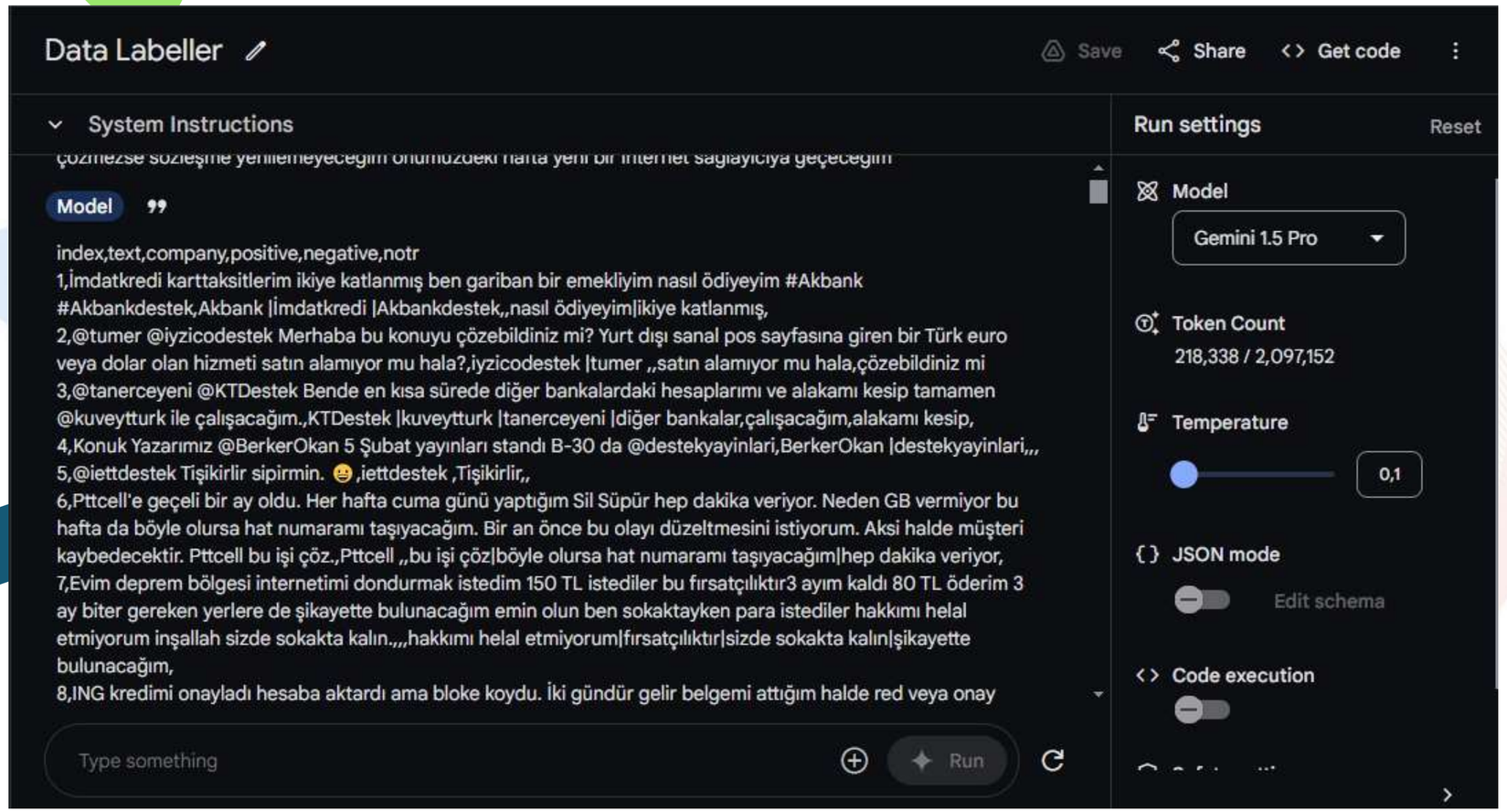
VERİ ETİKETLEME

Veri etiketleme sürecinde, çekilen veriler arasından 1000 tanesini özenle seçildi ve etiketlendirildi. İzlediğimiz adımlar şunlardır:

- **Veri Seçimi:** Çekilen veriler arasından, farklı durumları ve müşteri etkileşimlerini temsil eden 1000 veri seçildi.
- **Etiketleme Kriterlerinin Belirlenmesi:** Her veri için belirli etiketleme kriterleri oluşturduk. Örneğin, olumlu, olumsuz veya nötr duygular gibi kategoriler belirledik.
- **Manuel Etiketleme:** Seçilen 1000 veriyi, belirlenen kriterlere göre manuel olarak etiketledik. Bu süreçte, her bir veri dikkatle incelendi ve uygun etiketlerle işaretlendi.

Veri seti, makine öğrenimi modellerimizin eğitimi ve test edilmesi için kullanılmıştır. Müşteri geri bildirimlerini ve destek taleplerini daha iyi analiz edebilmemize olanak tanıyarak müşteri memnuniyetini artırmaya yönelik stratejiler geliştirmemize yardımcı olacaktır.

VERİ OTOMASYONU



Temparature	Score
0	33
0.1	34
0.25	31
0.50	25
0.75	25
1	27
1.25	21
1.75	25
2	22

Normalde çok uzun sürecek olan veri etiketleme işlemini yaptığımız model ile beraber otomasyonlaştırarak zamandan tasarruf edilmiştir.

Negatif Kategorisinde Geçen Favori Kelimeler

Nötr Kategorisinde Geçen Favori Kelimeler

Hem Pozitif Hem Negatif Kategorisinde Geçen Favori Kelimeler

Pozitif Kategorisinde Geçen Favori Kelimeler

MODELLENDİRME

Modellendirme aşamasında aşağıda yer alan birçok derin öğrenme modeline takımımız tarafından oluşturulan buna dayalı “Generative AI” yardımıyla üretilmiş sentetik verilerle ince ayar (finetune) yapıldı. Elde edilen sonuçlara bakılarak en yüksek doğruluk ve F1 skoruna sahip olan modelin kullanılması tercih edildi. Tercihimiz BERTürk oldu.

- BERTürk-Cased (128 binlik versiyonu)

  Accuracy: 91.6750
- Long Short-Term Memory

  Accuracy: 89.4700
- BERTürk-Uncased (base versiyonu)

  Accuracy: 90.5800

EĞİTİM VE DOĞRULAMA SÜRECİ

[5396/5396 11:48, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Accuracy	F1
1	0.213100	0.200786	0.935380	0.934999
2	0.181700	0.200490	0.938222	0.938016

```
comment = "Vodafone bazen adamı bezdiriyor ama güzel firma. Turkcell ise her daim mükemmel."

sentence = Sentence(comment)

ner_recognizer.predict(sentence)
```

```
[{'entity': 'Vodafone', 'sentiment': 'Pozitif'}, {'entity': 'Vodafone', 'sentiment': 'Negatif'}, {'entity': 'Turkcell', 'sentiment': 'Pozitif'}]
```

TEST SONUÇLARI

Tokenizer do_lower_case	True
Accuracy	90.5
F1 Score	90.3

Tokenizer do_lower_case	False
Accuracy	90.6
F1 Score	90.4

early_stopping_callback load_best_model_at_end metric_for_best_model	True
Accuracy	90.0
F1 Score	89.6

early_stopping_callback load_best_model_at_end metric_for_best_model	False
Accuracy	90.5
F1 Score	90.3

Model	Uncased
Accuracy	83.2
F1 Score	82.7

Model	Cased
Accuracy	0.832
F1 Score	0.827

TEST SONUÇLARI

Learning Rate	1e - 05
Accuracy	90.4
F1 Score	90.1

Learning Rate	1e - 06
Accuracy	90.0
F1 Score	89.8

Learning Rate	2e - 05
Accuracy	89.4
F1 Score	88.7

Schedular	linear
Max Acc	90.7
Max F1	90.6

Schedular	constant_with_warmup
Accuracy	90.2
F1 Score	90.1

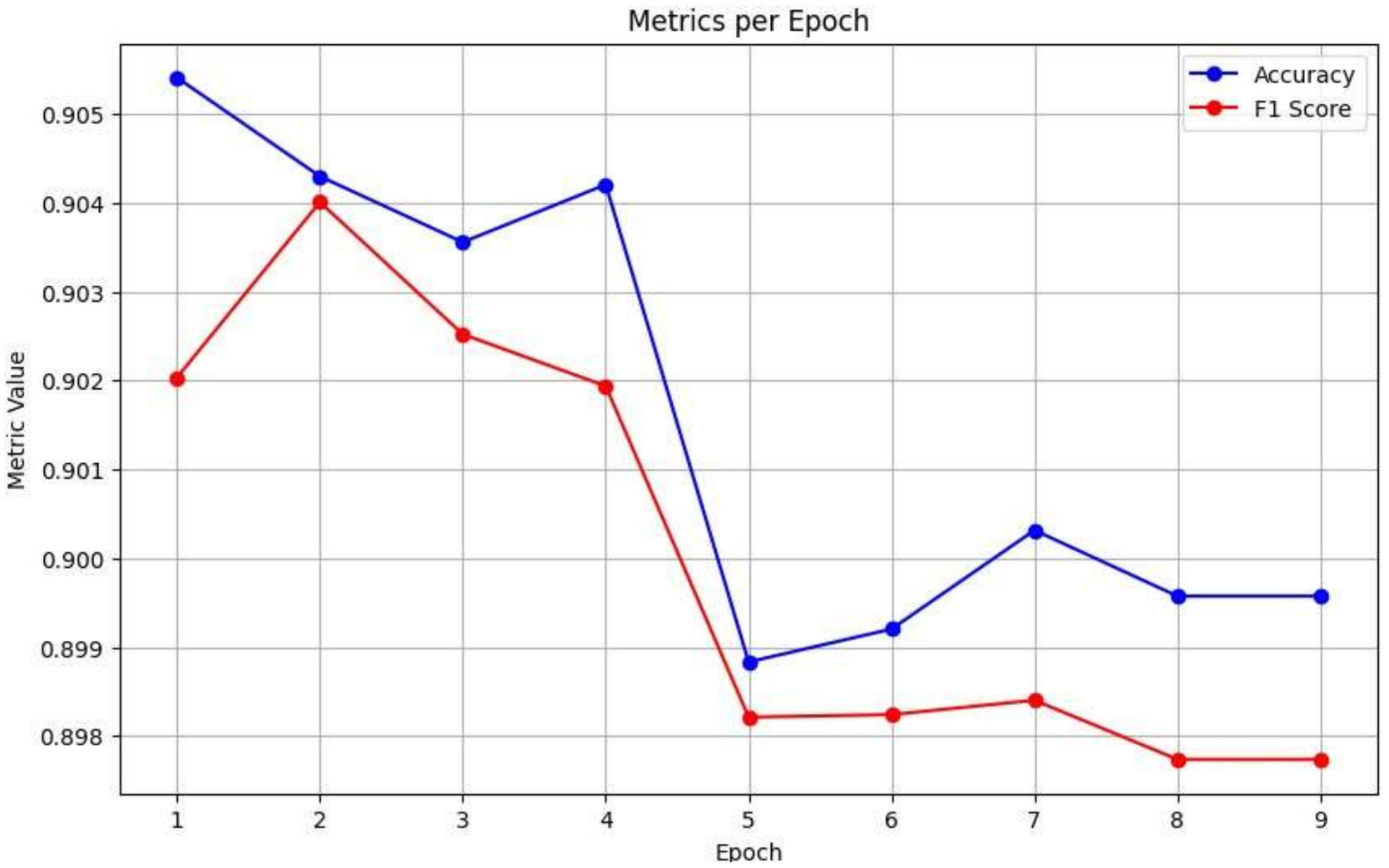
Schedular	inverse_sqrt
Accuracy	89.4
F1 Score	89.3

Optimizer	AdamW
Max Acc	90.7
Max F1	90.6

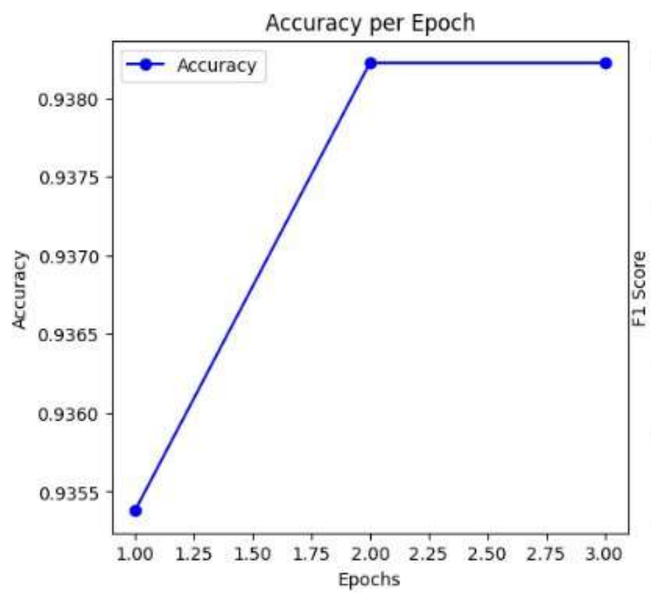
Optimizer	SGD
Accuracy	75.9
F1 Score	68.9

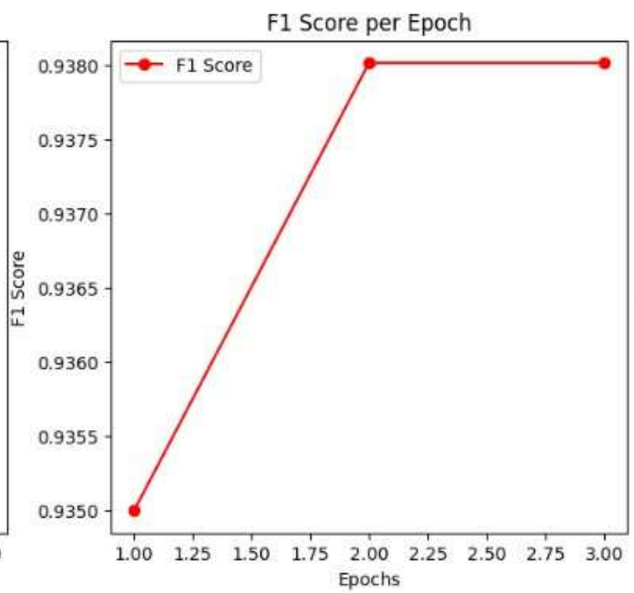
Optimizer	RMSprop
Accuracy	90.4
F1 Score	90.4

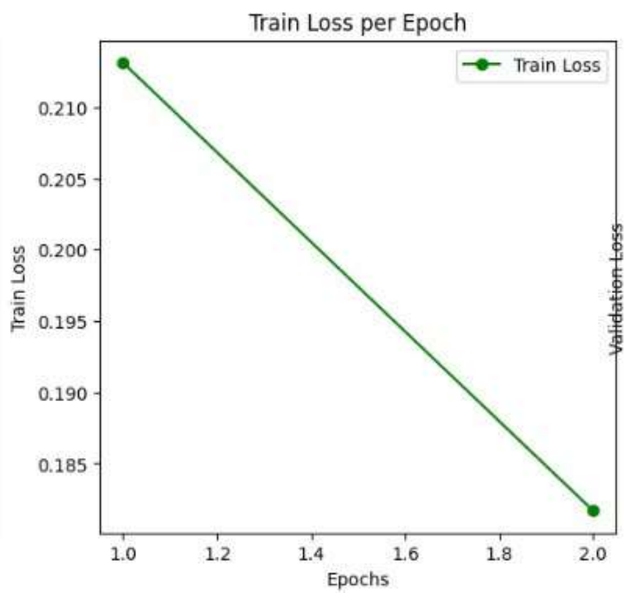
MODEL TESTLERİ

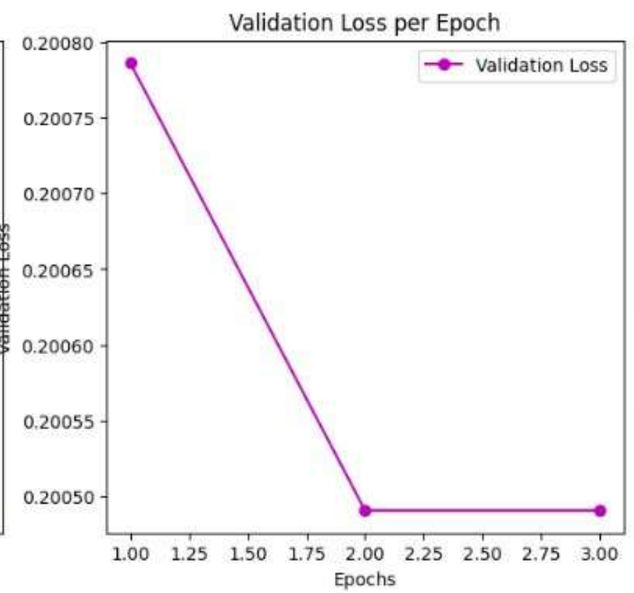


MODEL TESTLERİ

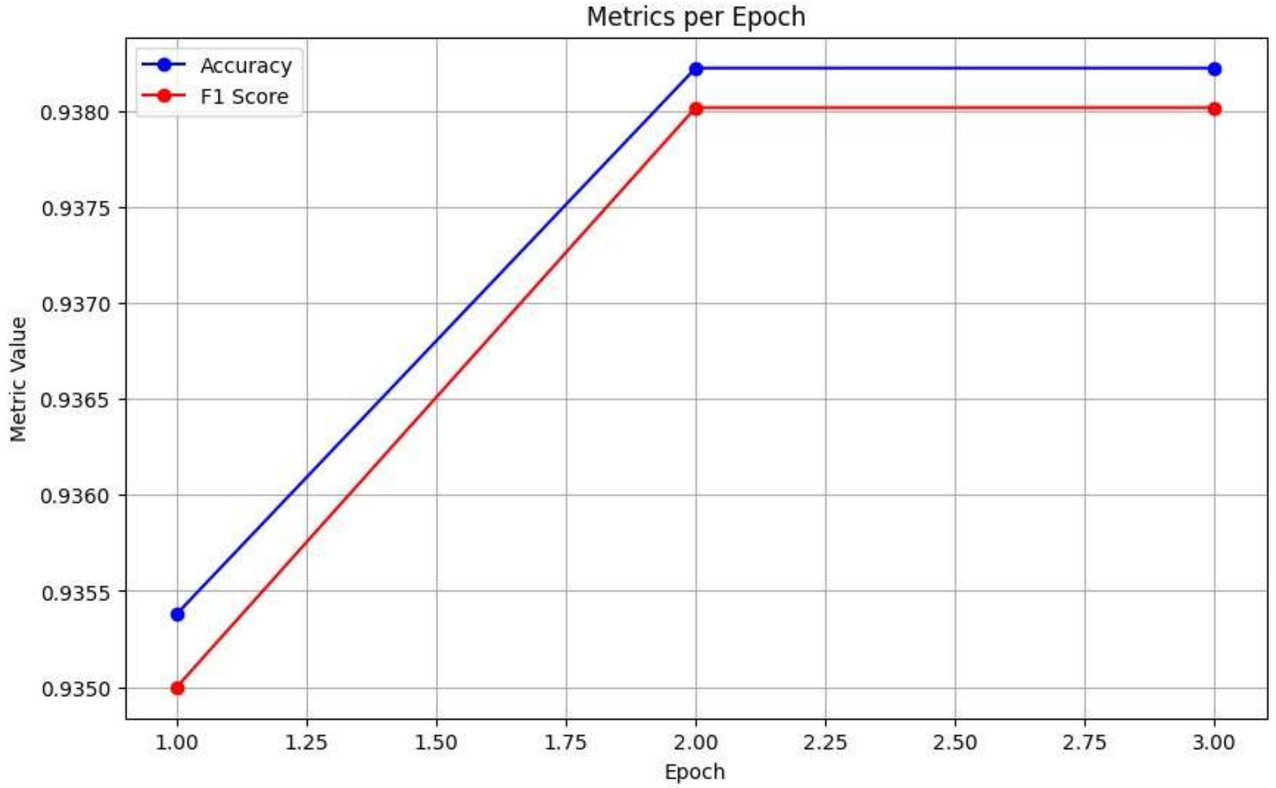


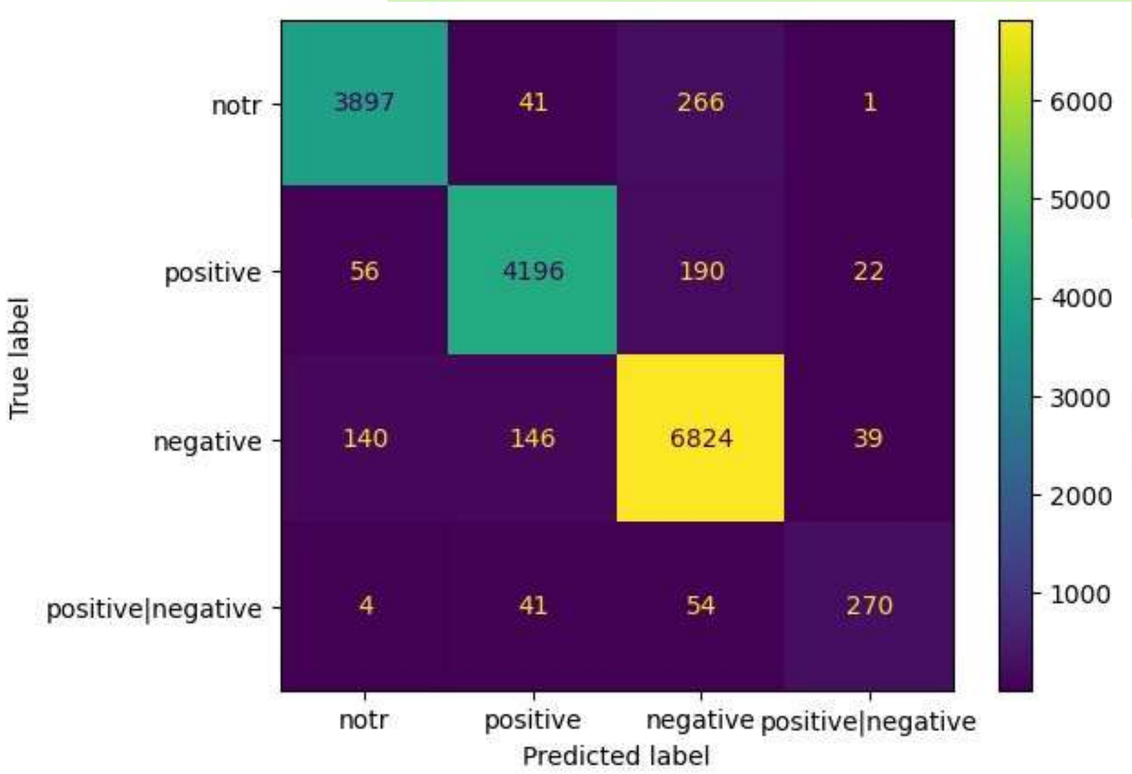




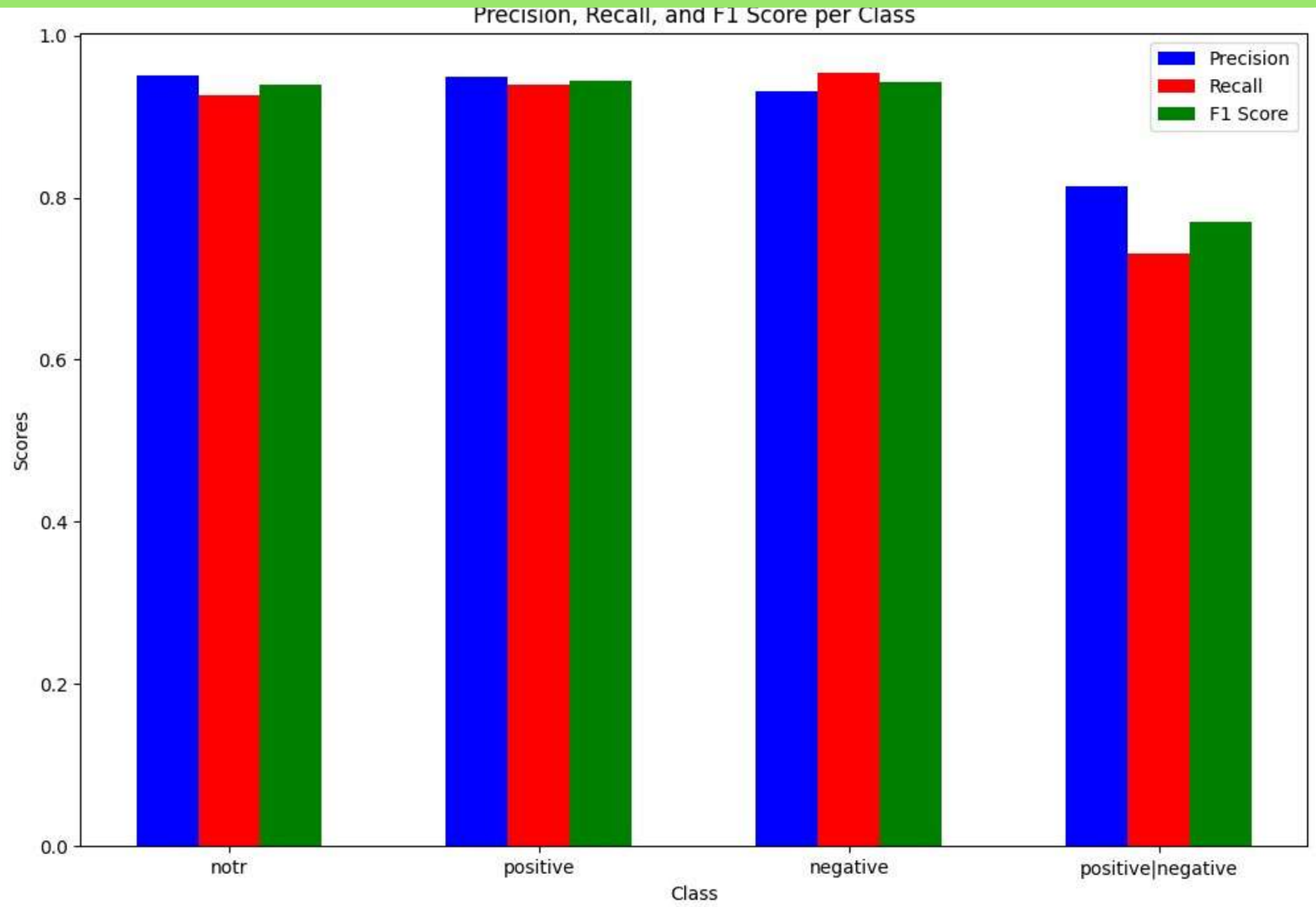


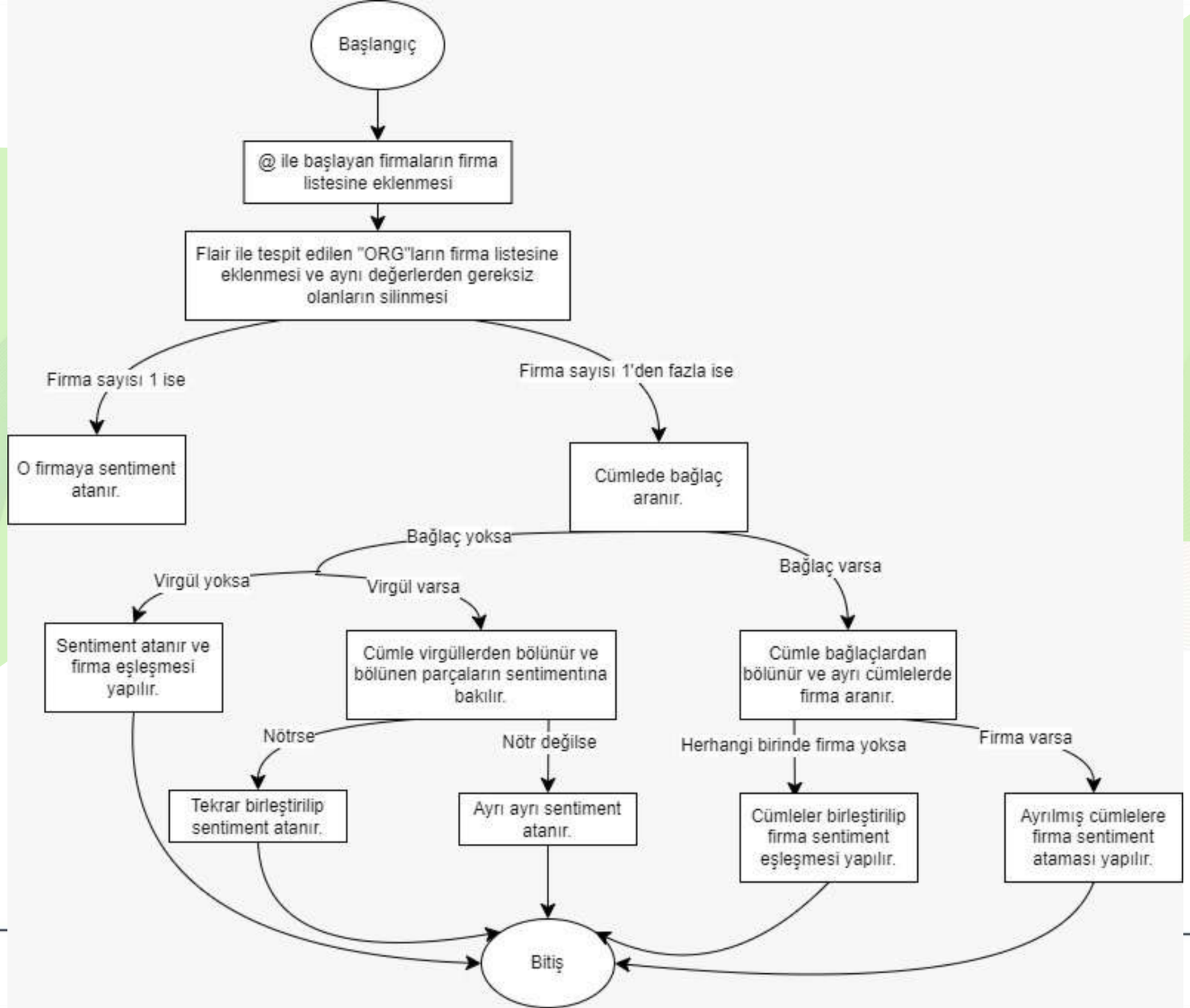
MODEL TESTLERİ

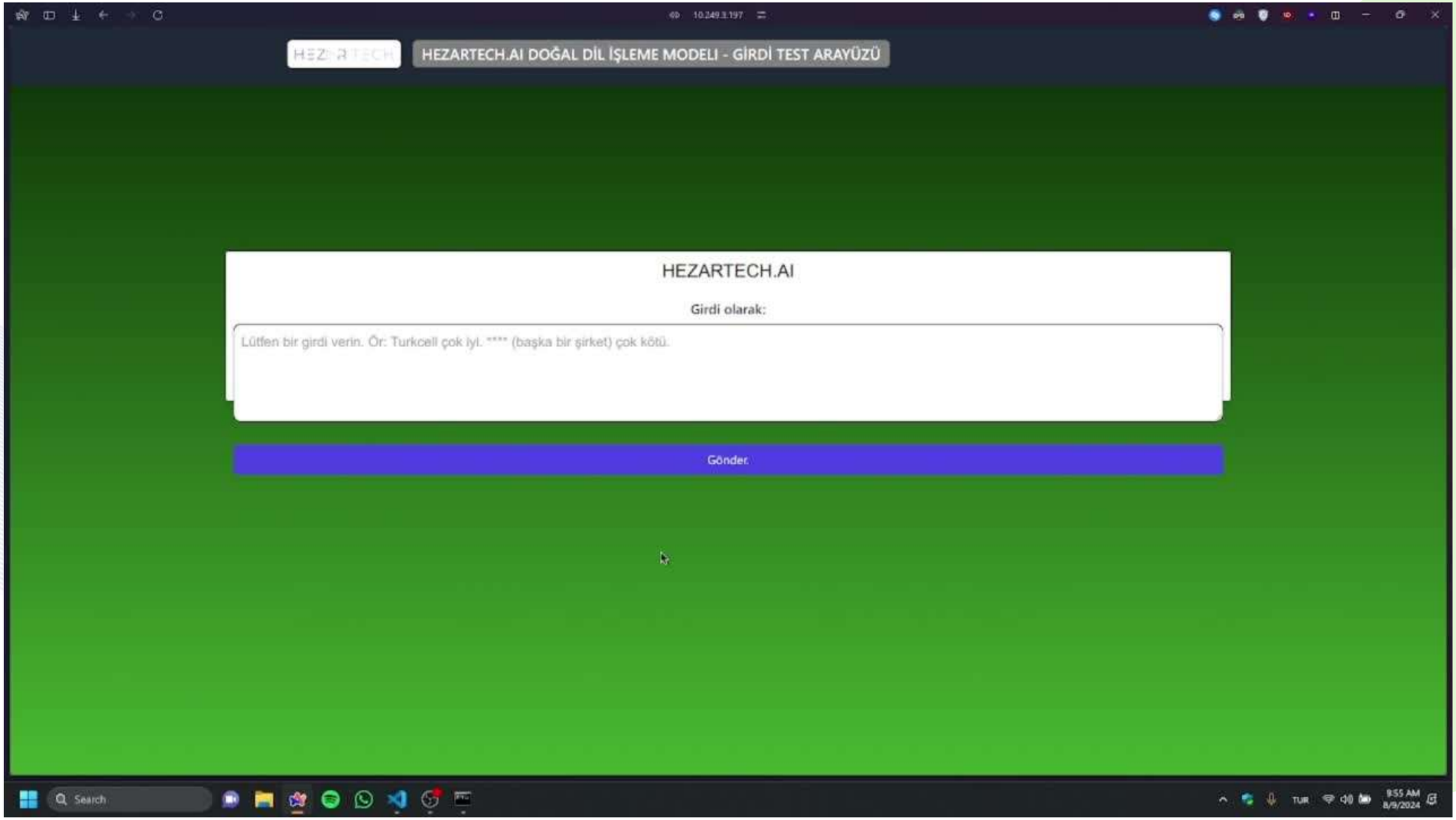




MODEL TESTLERİ




Başlangıç
@ ile başlayan firmaların firma listesine eklenmesi
Flair ile tespit edilen "ORG"ların firma listesine eklenmesi ve aynı değerlerden gereksiz olanların silinmesi
Firma sayısı 1 ise
Firma sayısı 1'den fazla ise
O firmaya sentiment atanır.
Cümlede bağlaç aranır.
Bağlaç yoksa
Bağlaç varsa
Virgül yoksa
Virgül varsa
Sentiment atanır ve firma eşleşmesi yapılır.
Cümle virgüllerden bölünür ve bölünen parçaların sentimentına bakılır.
Cümle bağlaçlardan bölünür ve ayrı cümlelerde firma aranır.
Nötrse
Nötr değilse
Herhangi birinde firma yoksa
Firma varsa
Tekrar birleştirilip sentiment atanır.
Ayrı ayrı sentiment atanır.
Cümleler birleştirilip firma sentiment eşleşmesi yapılır.
Ayrılmış cümlelere firma sentiment ataması yapılır.
Bitiş


HEZARTECH
HEZARTECH.AI DOĞAL DİL İŞLEME MODELI - GİRDİ TEST ARAYÜZÜ
HEZARTECH.AI
Girdi olarak:
Lütfen bir girdi verin. Ör: Turkcell çok iyi. **** (başka bir şirket) çok kötü.
Gönder.
Search
TUR
8:55 AM
8/9/2024

HEZARTECH

TEŞEKKÜRLER

8-9 AĞUSTOS 2024

BİLİŞİM VADİSİ

TEKNOFEST
HAVACILIK, UZAY VE TEKNOLOJİ FESTİVALİ