**Çalışma 1 Raporu:**

## 1. Giriş

Bu çalışmada, iki farklı veri seti üzerinde **duygu analizi (sentiment analysis)** gerçekleştirilmiştir:

1. **Winvoker yorumları (Türkçe)**
2. **Airline Tweets (İngilizce)**

Çalışma boyunca izlenen adımlar:

* Veri toplama ve yükleme
* Veri ön işleme (temizleme, lemmatization, stop-word temizleme, zemberek, NLTK)
* Vektörleştirme (BERT embedding)
* Model eğitimi (MLP, SVM, Logistic Regression, RF)
* Performans değerlendirmesi ve grafikler

## 2. Winvoker Dataseti (Türkçe)

### 2.1 Veri Toplama ve Yükleme

Winvoker dataset’i **HuggingFace datasets kütüphanesi** üzerinden yüklendi:

 train\_df ve test\_df olarak ikiye ayrılmıştır.

 text → Kullanıcı yorumu

 label → Yorumun duygu etiketi (pozitif, negatif, nötr)

### 2.2 Ön İşleme

Türkçe metinler üzerinde **Zemberek** kütüphanesi ile lemmatization yapılmıştır. Ayrıca:

1. Küçük harfe çevirme
2. Noktalama ve özel karakter temizleme
3. Tokenizasyon
4. Stop-word çıkarma (opsiyonel, Türkçe stop-word listesi ile)
5. İşlenmiş veriler joblib ile kaydedildi:

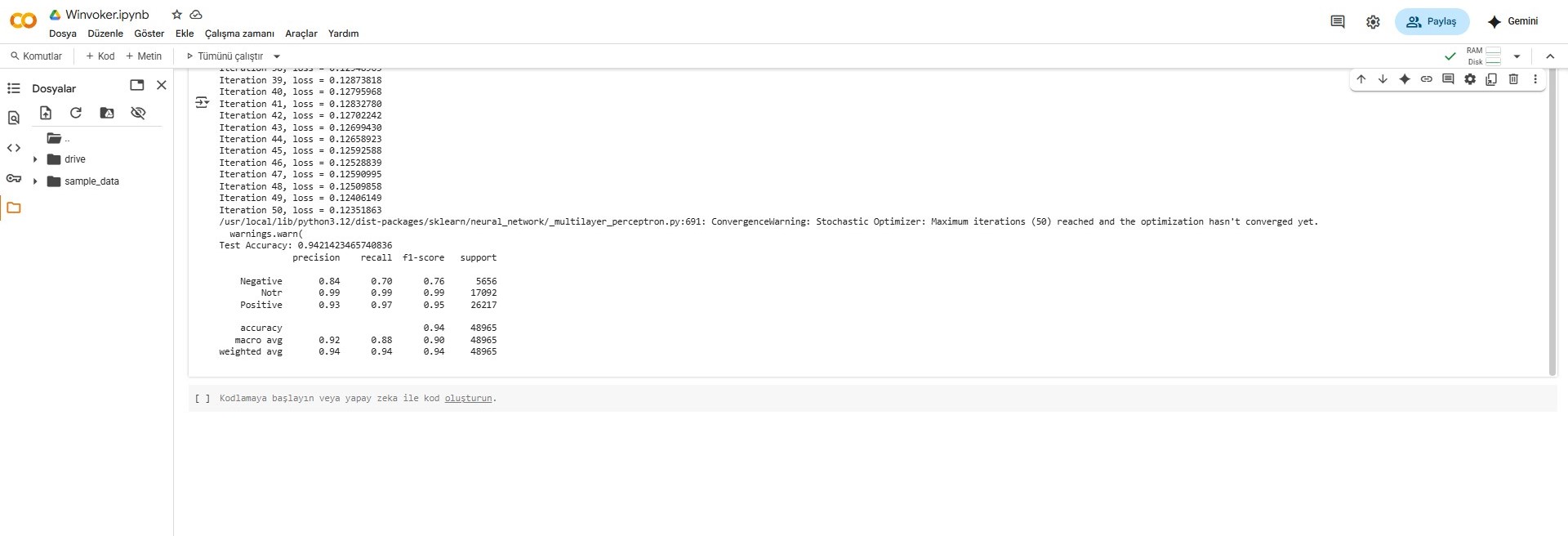
### 2.3 Vektörleştirme (BERT Embedding)

**BERT modeli (dbmdz/bert-base-turkish-cased)** kullanılarak metinler 768 boyutlu vektörlere dönüştürüldü.

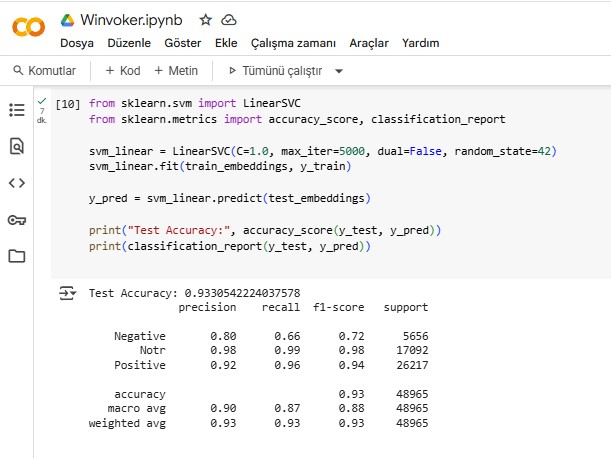
### 2.4 Model Eğitimi

Winvoker verisi için çeşitli modeller kullanıldı:

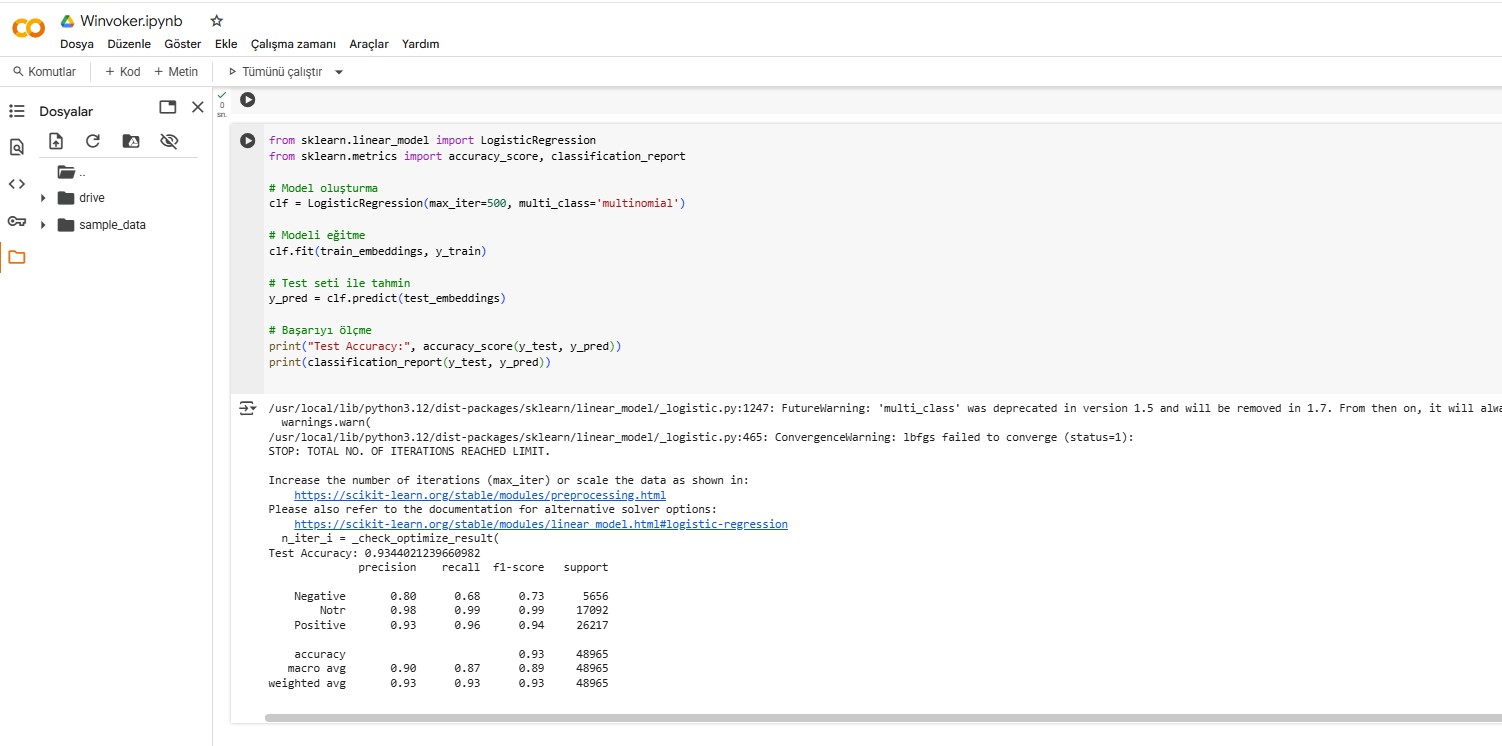
#### MLPClassifier



* Linear SVM



* Logistic Regression



## 3. Airline Tweets Dataseti (İngilizce)

### 3.1 Veri Toplama ve Yükleme

 text → Tweet metni

 airline\_sentiment → Duygu etiketi (positive, negative, neutral)

### 3.2 Ön İşleme

NLTK kullanılarak:

* Küçük harfe çevirme
* URL, mention ve hashtag temizleme
* Noktalama ve sayı temizleme
* Tokenizasyon
* Stop-word temizleme
* Lemmatization

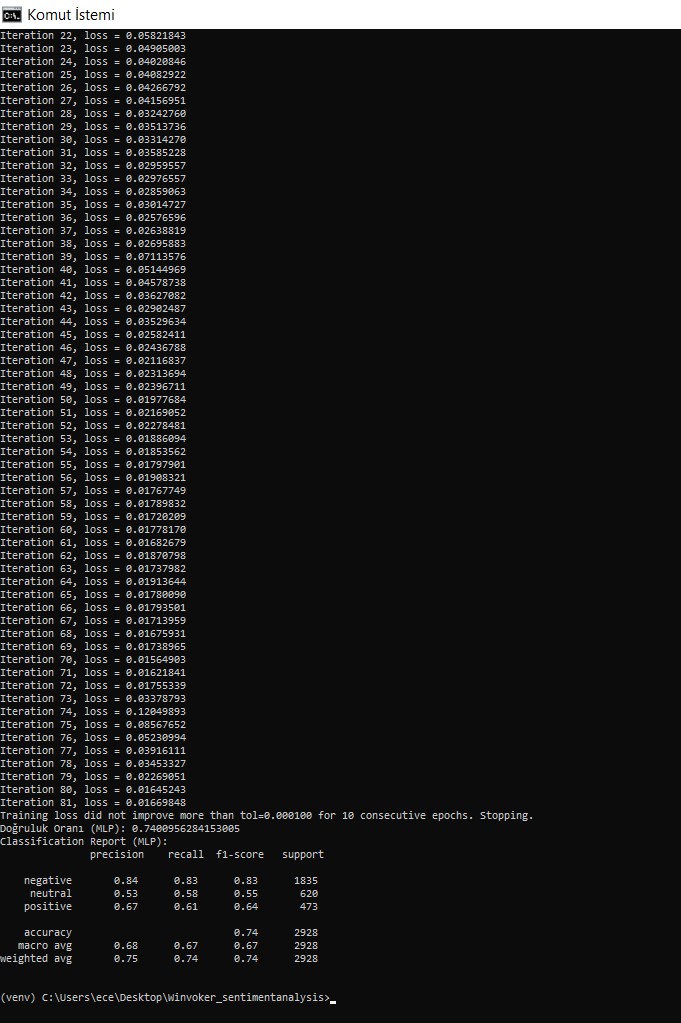
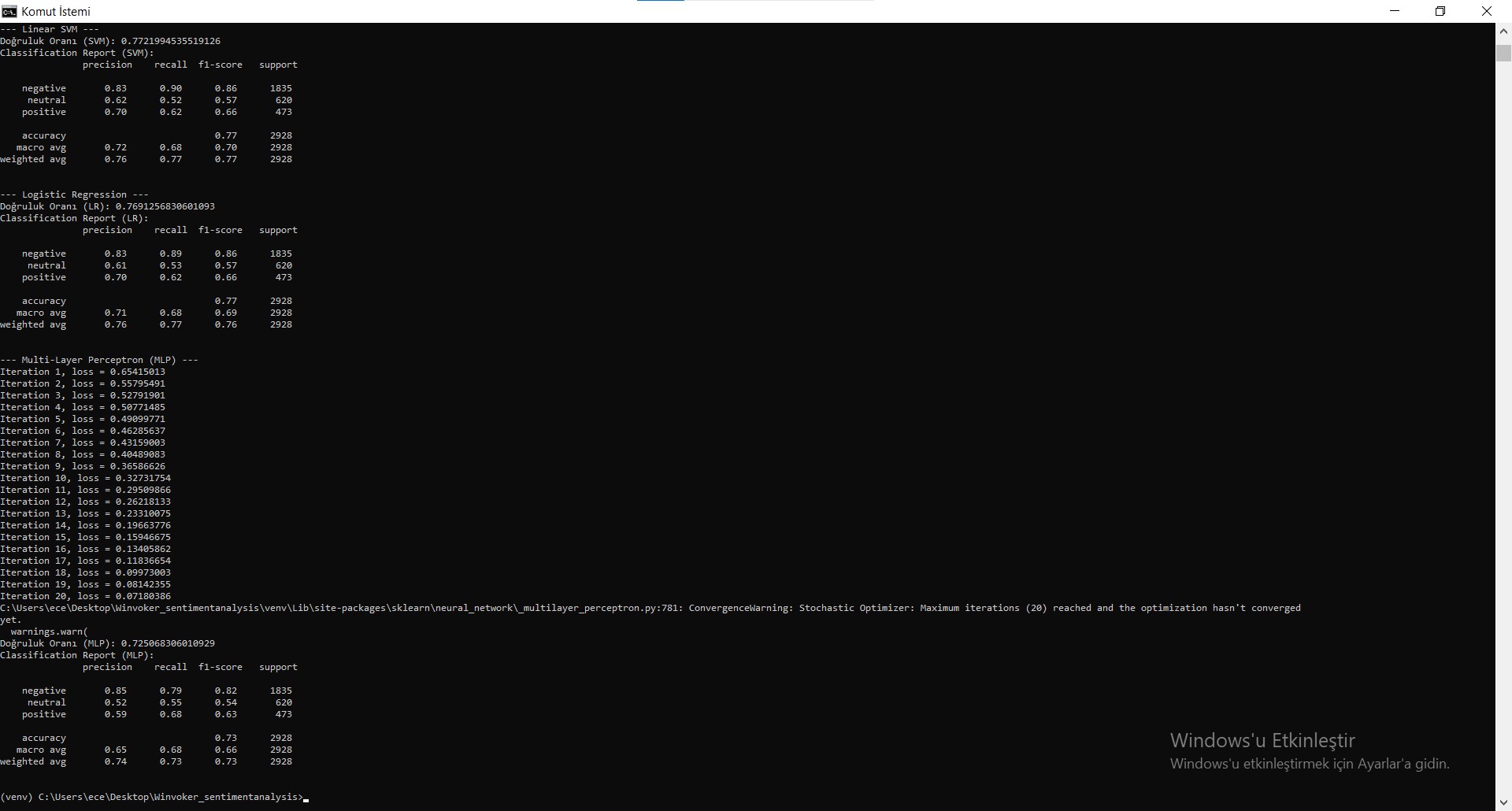
### 3.3 Vektörleştirme (BERT Embedding)

### 3.4 Model Eğitimi

Train-test bölme ve çeşitli modeller kullanıldı:

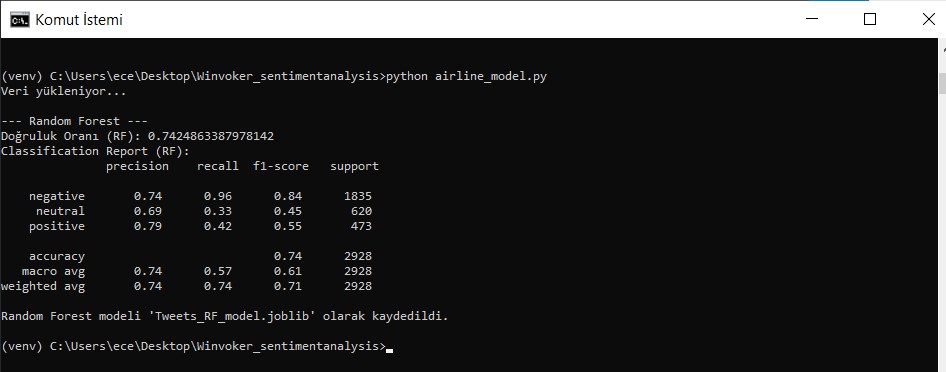
* Linear SVM
* Logistic Regression

#### MLPClassifier



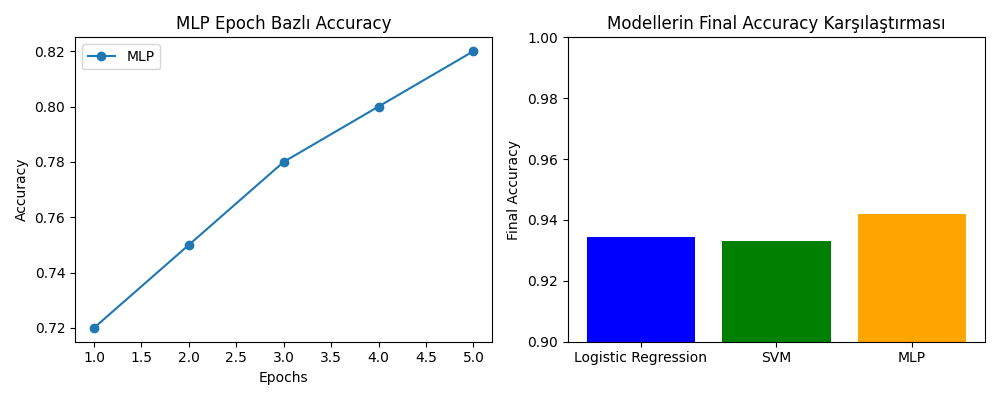
(MLP iteration artırıldı.)

* Random Forest



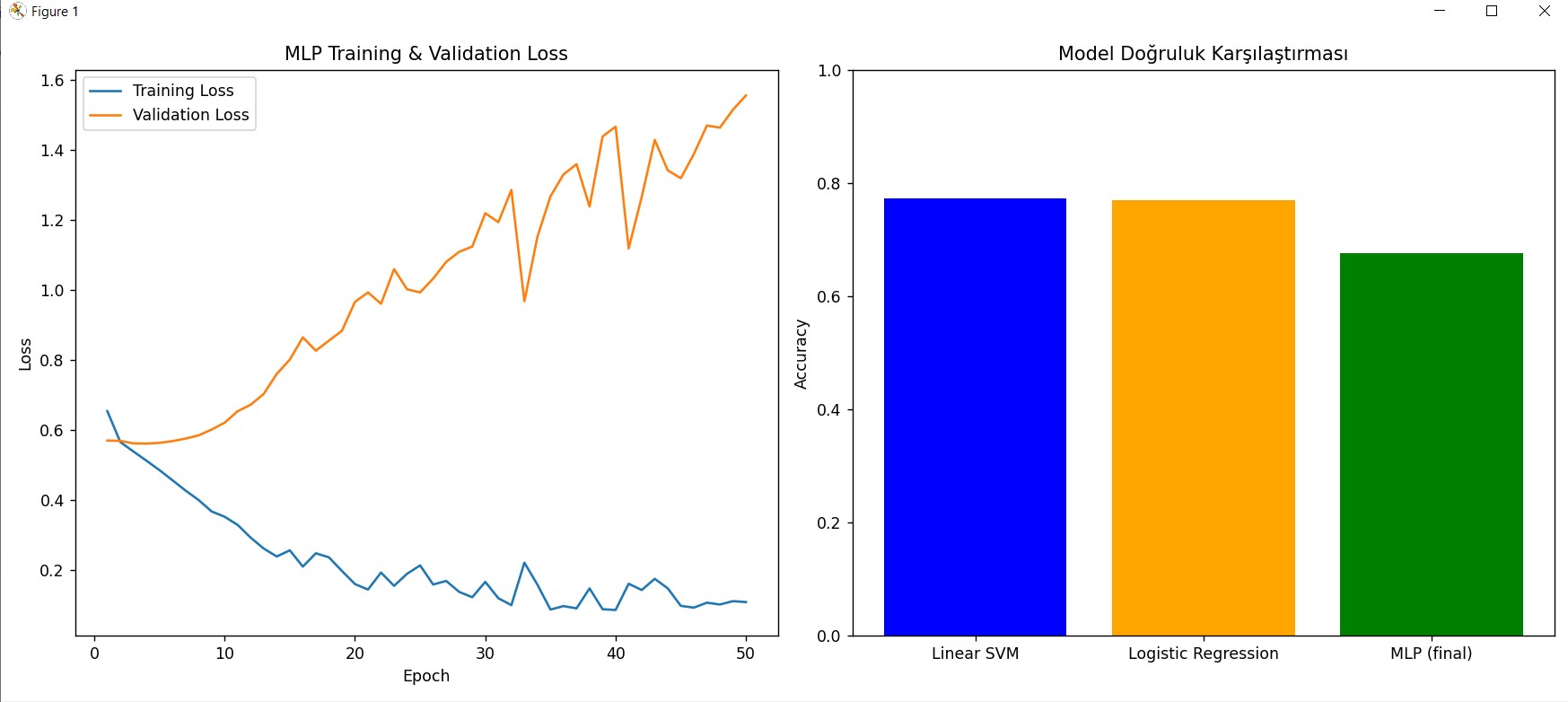
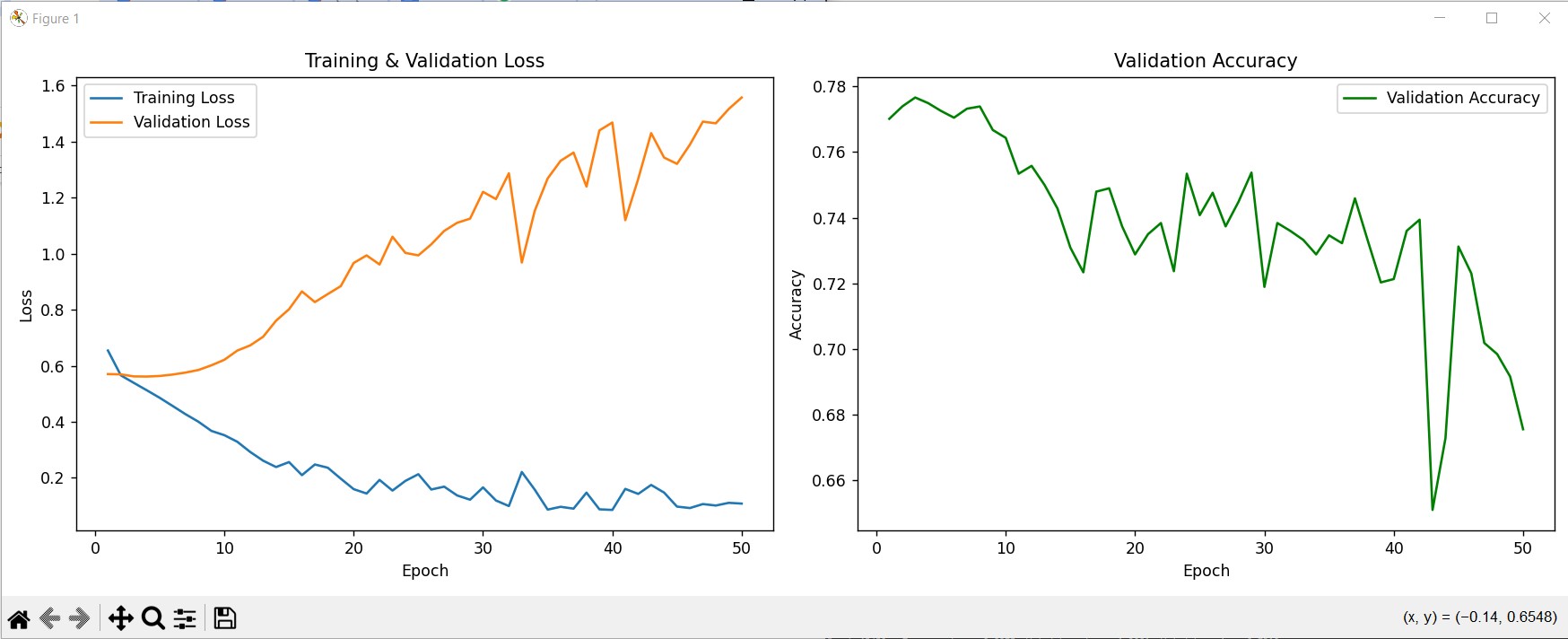
### 3.5 Grafikler

**Winvoker:**



**(Vs Code üzerinden)**

**Airline tweet:**



**(Vs Code üzerinden)**

## 4. Sonuç ve Değerlendirme

Bu çalışmada, farklı veri setleri üzerinde duygu analizi gerçekleştirilmiş ve modellerin performansları karşılaştırılmıştır. Analizler hem ön işleme ve vektörleştirme aşamalarını hem de modelleme sonuçlarını kapsamaktadır.

**Winvoker Dataseti Analizi:**

* Winvoker veri seti üzerinde Zemberek Türkçe doğal dil işleme kütüphanesi kullanılarak ön işleme yapılmış ve BERT tabanlı embedding ile metinler vektörleştirilmiştir.
* Üç farklı model uygulanmıştır: Logistic Regression (LR), Support Vector Machine (SVM) ve Multi-Layer Perceptron (MLP).
* Test doğrulukları:
  + LR: %93.44
  + SVM: %93.31
  + MLP: %94.21
* **Performans değerlendirmesi:**
  + MLP, özellikle pozitif sınıfı yüksek doğruluk ve F1-score ile tahmin edebilmiş ve genel doğrulukta diğer modellere göre biraz daha iyi performans göstermiştir.
  + LR ve SVM benzer performans göstermiştir; ancak negatif sınıfta recall değerleri düşük kalmıştır (%66–%68).
  + Model çıktılarında convergence uyarıları görülmüştür; bu, daha fazla iterasyon veya veri ölçeklendirme ile iyileştirilebilir.

**Airline Tweet Dataseti Analizi:**

* Airline veri seti üzerinde NLTK kullanılarak ön işleme yapılmış ve BERT embedding ile metinler vektörleştirilmiştir.
* Dört farklı model uygulanmıştır: LR, SVM, MLP ve Random Forest (RF).
* Test doğrulukları:
  + LR: %76.91
  + SVM: %77.22
  + MLP: %74.01( iteration artırıldı önceki denemede %72 değerindeydi.)
  + RF: %74.25
* **Performans değerlendirmesi:**
  + SVM ve LR, genel doğrulukta benzer sonuçlar vermiş ve nötr sınıf için düşük recall (%52–%53) göstermiştir.
  + MLP ve RF, negatif sınıf için yüksek doğruluk elde etmiş ancak pozitif ve nötr sınıflarda performans düşmüştür.
  + RF modeli, özellikle nötr ve pozitif sınıfın dengesizliği nedeniyle macro F1-score açısından daha düşük bir performans göstermiştir.

**Genel Çıkarımlar:**

* Winvoker veri seti, daha büyük ve dengeli olduğundan modeller genel olarak yüksek doğruluk göstermiştir. MLP en iyi performansı sağlamıştır.
* Airline veri seti, sınıflar arası dengesizlik nedeniyle doğruluk ve F1-score değerlerinde sınıfa bağlı farklılıklar ortaya çıkarmıştır.
* BERT embedding kullanımı, her iki veri setinde de metin temsilinde güçlü bir vektörleşme sağlamış ve klasik TF-IDF temelli yaklaşımlara göre daha yüksek performans göstermiştir.
* Görselleştirmeler, epoch başına kayıp ve doğruluk değişimlerini, modellerin karşılaştırmasını net biçimde ortaya koymuştur.
* Çalışmada ön işleme, vektörleştirme ve modelleme adımları adım adım uygulanmış, böylece farklı NLP ve makine öğrenmesi tekniklerinin veri üzerindeki etkisi gözlemlenmiştir.