

Tüketici Şikayeti Metinlerini Sınıflandırma

Bitirme Projesi Raporu

Esmâ Nur Beğbağâ
Yazılım Mühendisliği Bölümü

Ağustos 2026



softTo 3. DÖNEM 3. GRUP
BİLİŞİM GÜVENLİĞİ MİMARİSİ
BİTİRME PROJESİ RAPORU

Oracle WebLogic Server

Azra Öykü Ulukan
Mert Can Aydın
Ali Enes Çelikkalek

Ekim, 2025
softTo Yazılım Bilişim Akademisi, İstanbul

İçindekiler

1 GİRİŞ	3
1.1 Projenin Amacı	3
1.2 Projenin Önemi ve Kapsamı	3
1.3 Proje Kısıtları	3
2 MATERYAL VE YÖNTEM	3
2.1 Veri Seti (CFPB Consumer Complaints)	3
2.2 Doğal Dil İşleme (NLP) Ön İşleme Adımları	3
2.3 Özellik Çıkarımı ve Temsil Yöntemleri	4
2.4 Kullanılacak Modeller	4
2.5 Değerlendirme Metrikleri	4
3 SİSTEM MİMARİSİ VE UYGULAMA PLANI	4
4 BEKLENEN SONUÇLAR VE TARTIŞMA	5

1 GİRİŞ

1.1 Projenin Amacı

Bu projenin amacı, ABD Tüketici Mali Koruma Bürosu'na (CFPB) iletilen yapılandırılmamış gerçek müşteri şikayet metinlerini Doğal Dil İşleme (NLP) ve Makine Öğrenmesi yöntemleri kullanarak analiz etmek ve otomatik olarak ilgili finansal ürün/hizmet kategorilerine (Kredi Kartı, İpotek/Mortgage, Öğrenci Kredisi, Banka Hesapları vb.) yönlendiren çok sınıflı (multi-class) bir sınıflandırma sistemi geliştirmektir.

1.2 Projenin Önemi ve Kapsamı

Finansal kuruluşlar ve regülatör kurumlar her gün yüz binlerce şikayet almaktadır. Bu şikayetlerin manuel incelenmesi yüksek operasyonel maliyet ve zaman kaybı yaratır. Otomatik yönlendirme mekanizması yanıt sürelerini kısaltacak ve müşteri memnuniyetini artıracaktır. Proje; veri ön işleme, özellik çıkarımı, modelleme, performans karşılaştırması ve geliştirilen modelin bir API servisi olarak sunulması adımlarını kapsamaktadır.

1.3 Proje Kısıtları

- **Kısıt 1:** CFPB veri seti çok büyük olduğundan (milyonlarca satır), eğitim süreci yalnızca şikayet metni ("Consumer complaint narrative") alanı dolu olan filtrelenmiş bir alt küme (ör. 50.000 - 100.000 satır) üzerinde yürütülecektir.
- **Kısıt 2:** Metinler İngilizce dili ile sınırlıdır; çok dilli genişleme kapsam dışı bırakılmıştır.
- **Kısıt 3:** Donanım ve eğitim süresi sınırları nedeniyle modeller yerel GPU/bulut ortamı limitlerine göre optimize edilecektir.

2 MATERYAL VE YÖNTEM

2.1 Veri Seti (CFPB Consumer Complaints)

Projede CFPB'nin resmi açık veri portalından temin edilen tüketici şikayetleri veritabanı kullanılacaktır.

- **Girdi Özelliği:** Consumer complaint narrative (Müşteri şikayet metni)
- **Hedef Değişken:** Product (Finansal ürün sınıfı — ö. Credit card, Mortgages, Student loan, Debt collection)

2.2 Doğal Dil İşleme (NLP) Ön İşleme Adımları

- **Küçük Harfe Dönüştürme ve Noktalama Temizliği:** Metinler standartlaştırılır ve gürültüler temizlenir.
- **Stop-words Temizliği:** "the, is, at" gibi anlamsal ayırt ediciliği düşük sözcükler ayıklanır.
- **Lemmatizasyon / Stemming:** Sözcükler morfolojik kök formlarına indirgenir.

- **Anonimleştirilmiş İfadelerin Temizlenmesi:** Veri setinde kişisel verilerin gizlenmesi için yer alan “XXXX” şeklindeki maskelenmiş özel ifadeler kaldırılır.

2.3 Özellik Çıkarımı ve Temsil Yöntemleri

- **TF-IDF (Term Frequency - Inverse Document Frequency):** N-gram tabanlı istatistiksel kelime frekans matrisleri oluşturulur.
- **Transformer Tabanlı Embeddings (BERT / RoBERTa):** Cümle ve bağlam düzeyinde anlamsal vektör temsilleri çıkarılır.

2.4 Kullanılacak Modeller

1. **Temel Modeller (Baseline):** TF-IDF öznitelikleriyle eğitilen Lojistik Regresyon (Logistic Regression), Multinomial Naive Bayes ve Random Forest.
2. **Derin Öğrenme / Transformer Modeller:** İnce ayar (fine-tuning) yapılmış bert-base-uncased veya DistilBERT mimarisi.

2.5 Değerlendirme Metrikleri

Sınıflar arasındaki veri dengesizliğini (class imbalance) doğru analiz edebilmek için aşağıdaki metrikler kullanılacaktır:

- Accuracy (Doğruluk)
- Precision, Recall ve F1-Score (Macro & Weighted)
- Karmaşıklık Matrisi (Confusion Matrix)

3 SİSTEM MİMARİSİ VE UYGULAMA PLANI

- **Geliştirme Ortamı:** Python 3.10+, Jupyter Lab/VS Code, PyTorch, HuggingFace Transformers, Scikit-learn, Pandas, FastAPI.
- **Veri Temizleme & EDA:** Eksik verilerin tespiti, sınıf dağılım grafikleri, en sık geçen kelime/n-gram analizleri.
- **Model Eğitimi:** Veri %80 Eğitim, %20 Test olarak ayrılacak; Stratified K-Fold Cross Validation ile aşırı öğrenme (overfitting) kontrol edilecektir.
- **API Servisi:** Eğitilen en başarılı model FastAPI üzerinden bir /predict uç noktası ile canlıya alınarak rastgele girilen şikayet metnine kategori tahmini ve güven skoru üretecektir.
- **Test ve Doğrulama:** Gerçek zamanlı test senaryoları ve API performans testleri (yanıt süresi, eşzamanlı istekler) yapılacaktır.

4 BEKLENEN SONUÇLAR VE TARTIŞMA

- TF-IDF tabanlı klasik makine öğrenmesi modellerine kıyasla Transformer (BERT) tabanlı mimarinin bağlamsal ilişkileri daha başarılı öğrenerek özellikle metin uzunluğu fazla olan şikayetlerde belirgin bir F1-skoru artışı sağlaması beklenmektedir.
- Elde edilen modelin FastAPI ile servis edilmesi, projenin sadece deneysel bir çalışma olarak kalmayıp gerçek kurumsal iş akışlarına entegre edilebilir bir çözüm sunduğunu gösterecektir.

KAYNAKÇA

1. Consumer Financial Protection Bureau (CFPB), *Consumer Complaint Database*, <https://www.consumerfinance.gov/data-research/consumer-complaints/>
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805.
3. Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825-2830.
4. Vaswani, A., et al. (2017). *Attention Is All You Need*. Advances in Neural Information Processing Systems (NeurIPS).