MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE METİNLERDE SORGUYA UYGUN KISIMLARIN TOKEN SEVİYESİNDE TESPİTİ İÇİN GEÇ ETKİLEŞİMLİ FÜZYON MODELİ

SALİH CAN TURAN

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı

DANIŞMAN

Doç. Dr. Kazım YILDIZ

EŞ-DANIŞMAN

Dr. Öğr. Üyesi Ali SARIKAŞ

İSTANBUL, 2025



MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ



TÜRKÇE METİNLERDE SORGUYA UYGUN KISIMLARIN TOKEN SEVİYESİNDE TESPİTİ İÇİN GEÇ ETKİLEŞİMLİ FÜZYON MODELİ

SALİH CAN TURAN

523622002

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

DANIŞMAN

Doç. Dr. Kazım YILDIZ

EŞ-DANIŞMAN

Dr. Öğr. Üyesi Ali SARIKAŞ

İSTANBUL, 2025

ÖNSÖZ

Bu tez çalışması, Türkçe doğal dil işleme alanında hibrit etkileşim modelleri üzerine yapılan araştırmaların bir ürünüdür.

Tez çalışmamın her aşamasında hiçbir zaman desteğini esirgemeyen, her fırsatta yardımıma koşan değerli danışmanım Sn. Doç. Dr. Kazım YILDIZ'a,

Tez sürecimin her aşamasında değerli katkılarını ve desteklerini esirgemeyen eş-danışmanım Sn. Dr. Öğr. Üyesi Ali SARIKAŞ'a,

Bu çalışmanın gerçekleştirilmesinde gerekli olan akademik alt yapıyı sağlayan Marmara Üniversitesi Fen Bilimleri Enstitüsü ve Bilgisayar Mühendisliği Anabilim Dalı öğretim üyelerine

Teşekkürlerimi sunarım.

Ağustos 2024 Salih Can TURAN

MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Marmara Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Öğrencisi **Salih Can TURAN**'ın **"Türkçe Metinlerde Sorguya Uygun Kısımların Token Seviyesinde Tespiti için Geç Etkileşimli Füzyon Modeli"** başlıklı tez çalışması, ../../.... tarihinde savunulmuş ve jüri üyeleri tarafından başarılı bulunmuştur.

Jüri Üyeleri

Doç. Dr. Kazım YILDIZ (Danışman)
Marmara Üniversitesi ..... (İmza)

Dr. Öğr. Üyesi Timur İNAN (Üye)
Marmara Üniversitesi ..... (İmza)

Dr. Öğr. Üyesi Vecdi Emre LEVENT (Üye)
Fenerbahçe Üniversitesi ..... (İmza)

ONAY

Marmara Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun ../../.... tarih ve .... sayılı kararı ile **Salih Can TURAN**'ın Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Programında Yüksek Lisans derecesi alması onanmıştır.

Prof. Dr. Bülent MERTOĞLU
Fen Bilimleri Enstitüsü Müdürü

İÇİNDEKİLER

ÖNSÖZ i

ÖZET ix

ABSTRACT xi

KISALTMALAR xiii

SEMBOLLER xv

1 GİRİŞ 1

1.1 Problemin Tanımı . . . 1
1.2 Amaç ve Kapsam . . . 1
1.3 Araştırmanın Önemi ve Literatüre Katkısı . . . 2
1.4 Araştırma Soruları ve Hipotezler . . . 2
1.5 Araştırma Tanımları . . . 3
1.6 Sınırlılıklar ve Varsayımlar . . . 4
1.6.1 Sınırlılıklar . . . 4
1.6.2 Varsayımlar . . . 4

2 KAVRAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR 5

2.1 Etkileşim Mimarilerinde Temel Paradigmalar . . . 5
2.1.1 Çift Kodlayıcı Mimarileri: Verimlilik Odaklı Yaklaşım . . . 5
2.1.2 Çapraz Kodlayıcı Mimarileri: Başarım Odaklı Yaklaşım . . . 6
2.1.3 Geç Etkileşimli Mimariler . . . 6
2.2 Transformer Mimarisi ve Dikkat Mekanizması . . . 7
2.2.1 Transformer'ın Temel Bileşenleri . . . 7
2.2.2 Transformer'ın Doğal Dil İşlemedeki Etkisi . . . 8
2.2.3 Aralık Çıkarımı ve Dikkat Mekanizması . . . 8
2.2.4 Çalışmamızla İlişkisi . . . 9
2.3 Büyük Dil Modellerinde Alaka Kavramının Yapısal Analizi . . . 9
2.4 Gelişmiş Etkileşim Mimarileri ve Optimizasyon Stratejileri . . . 9
2.4.1 Erken ve Geç Etkileşim Mekanizmaları . . . 10
2.4.2 Liste Bazlı Optimizasyonlar ve CROSS-JEM . . . 10
2.4.3 Sıralama Odaklı Kayıp Fonksiyonları . . . 10

2.5 Etkileşim Mimarilerinde Güncel Yaklaşımlar ve Alternatifler 11
2.5.1 Alternatif Girdi Temsilleri: Token-Free Yaklaşımlar 11
2.6 Araştırma Boşluğu ve Tezin Konumlandırılması 12
2.7 Türkçe Doğal Dil İşleme: Morfolojik Zenginlik ve Çözüm Arayışları . 13
2.7.1 Türkçe'nin Morfolojik Özellikleri ve NLP Zorlukları 13
2.7.2 Türkçe BERT Varyantları: Karşılaştırmalı Analiz 14
2.7.3 Türkçe Token Sınıflandırma ve NER: Ortak Zorluklar ve Çözümler 14
2.7.4 Türkçe Token Sınıflandırma Görevlerindeki Özel Zorluklar . . 15
2.7.5 Çözüm Yaklaşımları ve Gelecek Yönelimler 15
2.8 Modern Dil Modellerinin Temel Taşları 16
2.8.1 Ardışık Veri Modelleme ve Uzun Vadeli Bağımlılık Sorunu . . 16
2.8.2 Kelime Temsilleri: Word2Vec ve GloVe 16
2.9 Sıfır Örnekli Öğrenme Paradigmaları ve Varlık İlişkilendirme 17
2.9.1 Çoklu Görev Tabanlı Eğitim ve Sıfır Örnekli Genelleme 17
2.9.2 Çift Aşamalı Yaklaşımlar: Bi-encoder ve Cross-encoder 17
2.9.3 Üretim Tabanlı Yaklaşımlar: Üretimle Getirme 18
2.9.4 Sıfır Örnekli Sınıflandırmanın Yeniden Formüle Edilmesi . . . 18
2.9.5 Zero-Shot Token Etiketleme Paradigmaları 19
2.9.6 Üretken ve Ayırt Edici Yaklaşımlar 20
2.9.7 Çapraz Dilli Modellerde Açık Token Hizalama 20
2.9.8 Yoğun Belge Erişimi 21
2.9.9 Passage-Level Evidence ve Token-Level Evidence 21
2.9.10 Büyük Dil Modelleri ile Veri Zenginleştirme 21

3 MATERYAL VE YÖNTEM 23
3.1 Veri Seti ve Hazırlık Süreci 23
3.1.1 Veri Kaynağı ve LLM Destekli Zenginleştirme 23
3.1.2 Evren ve Örneklem: Endüstriyel Odaklı Veri Seti 26
3.1.3 Verilerin Toplanması 27
3.1.4 Veri Ön İşleme ve Tokenizasyon 27
3.1.5 Temel Dil Modeli Seçimi: Artiwise ModernBERT 27
3.2 Deneysel Altyapı: ZSCAT Platformu 28
3.3 Model Mimarisi: Çift Kule Kodlayıcı ve Geç Etkileşimli Füzyon . . . 33
3.3.1 Sınıflandırma ve Çıktı Katmanları 37
3.3.2 Mimari Seçiminin Gerekçelendirilmesi 37
3.4 Eğitim Stratejisi: Sınıf Dengesizliği ve Odaklanmış Öğrenme 38
3.4.1 Eğitim Altyapısı ve Teknik Detaylar 39

3.4.2 Gelecek Çalışmalar İçin Potansiyel Bir İyileştirme: Kendi Kendine Öğrenme . . . 40
3.5 Etik Değerlendirmeler . . . 41
3.6 Referans Modeller . . . 41
3.6.1 Sıralı Kelime Grubu Eşleştirme . . . 41
3.6.2 Anlamsal Arama . . . 42
3.6.3 Model Mimarilerinin Verimlilik ve Hesaplama Maliyeti Analizi 42
3.7 Değerlendirme Metrikleri . . . 43
3.8 Detaylı Model ve Eğitim Parametreleri . . . 44
3.8.1 Temel Model Parametreleri . . . 44
3.8.2 Füzyon Katmanı Parametreleri . . . 45
3.8.3 Veri Seti Örnekleri . . . 45
3.9 LLM Destekli Veri Zenginleştirme . . . 46

4 BULGULAR VE TARTIŞMA 47
4.1 Deneysel Sonuçlar ve Analiz . . . 47
4.2 Hata Analizi ve Sınırlılıklar . . . 48
4.3 Model Mimarilerinin Verimlilik ve Hesaplama Maliyeti Analizi . . . 50

5 SONUÇ VE ÖNERİLER 51
5.1 Temel Bulgular . . . 51
5.2 Bulguların Değerlendirilmesi ve Tartışma . . . 51
5.3 Uygulayıcılar İçin Öneriler . . . 52
5.4 Gelecek Çalışmalar ve Araştırma Yönelimleri . . . 53

ÖZET

TÜRKÇE METİNLERDE SORGUYA UYGUN KISIMLARIN TOKEN SEVİYESİNDE TESPİTİ İÇİN GEÇ ETKİLEŞİMLİ FÜZYON MODELİ

Bu tez çalışmasında, Türkçe metinlerde kullanıcı sorgularına uygun olan kısımların otomatik olarak tespit edilmesi ve vurgulanması problemi ele alınmıştır. Çalışmanın temel amacı, az sayıda etiketli veri ile etkili bir token seviyesi sınıflandırma modeli geliştirerek, mevcut literatürdeki veri kıtlığı ve verimlilik sorunlarına çözüm sunmaktır.

Geleneksel bilgi erişim sistemlerinin belge seviyesinde çalışmasının aksine, bu çalışma, metin içerisindeki token düzeyinde ayrıntılı bir analiz sunmaktadır. Bu yaklaşım, kullanıcı sorgularına tam olarak hangi ifadelerin karşılık geldiğini saptayarak daha hassas sonuçlar üretmeyi amaçlamaktadır. Türkçe gibi eklemeli dillerin morfolojik zenginliği, bu seviyedeki bir analizi özellikle zorlu kılmaktadır.

Literatür incelemesi sonucunda, mevcut çalışmaların İngilizce metinler üzerinde yoğunlaştığı ve Türkçe için yeterli kaynak bulunmadığı tespit edilmiştir. Mevcut çözümler genellikle ya yüksek hesaplama maliyeti gerektiren cross-encoder mimarileri ya da düşük performanslı bi-encoder yaklaşımları kullanmaktadır.

Önerilen yaklaşım, `artiwise-ai/modernbert-base-tr-uncased` temel modeli üzerine "Geç Etkileşimli Füzyon" adı verilen hibrit bir mimari kullanmaktadır. Bu mimari iki aşamadan oluşur: İlk aşamada sorgu ve metin birbirinden bağımsız olarak kodlanır, ikinci aşamada özel füzyon katmanı ile bu temsiller arasında zengin etkileşim kurulur. Bu yaklaşım, bi-encoder'ların verimliliğini korurken cross-encoder'ların yüksek performansını yakalamayı hedefler.

Sınıf dengesizliği problemini çözmek için Focal Loss tabanlı eğitim stratejisi benimsenmiştir. Bu yaklaşım, pozitif ve negatif token'lar arasındaki dengesizliği gidererek modelin zorlu örnekleri öğrenmesini sağlar.

Veri seti, internet üzerindeki çeşitli kaynaklardan derlenen 500 bin adet Türkçe metnin, bu tezde geliştirilen LLM-destekli yarı denetimli bir yaklaşımla etiketlenmesiyle oluşturulmuştur. Bu süreç sonucunda, yaklaşık 2 milyon adet etiketli sorgu-metin çifti elde edilmiştir. Veri seti haber makaleleri, teknik blog yazıları, ürün yorumları ve forum yazıları gibi çeşitli kaynaklardan derlenmiştir.

Model eğitimi, 2 adet NVIDIA H100 GPU üzerinde 10 epoch boyunca gerçekleştirilmiştir. Doğrulama seti (validation set) üzerindeki başarımın 2. epoch'ta tepe noktasına ulaştığı gözlemlenmiş ve nihai değerlendirmeler için bu en iyi kontrol noktası seçilmiştir. Eğitim sürecinde batch size 32, learning rate 2e-5 ve AdamW

optimizer kullanılmıştır.

Deneysel sonuçlar, önerilen modelin en iyi kontrol noktasında (epoch 2) ulaştığı 0.557 F1-skoru ile temel 'Keyword Matching' (0.065) ve 'Klasik Bi-Encoder' (0.481) modellerine göre belirgin bir üstünlük sağladığını göstermektedir. Ayrıca, modelin performansı, SOTA bir referans model olan 'GLiNER''ın (0.580) performansına oldukça yakındır.

Verimlilik açısından, önerilen model, mimari olarak kendisinden daha karmaşık olan GLiNER modeline kıyasla daha az parametre kullanırken, çıkarım hızı olarak da avantajlar sunmaktadır. Bu durum, modelin pratik uygulamalardaki kullanım potansiyelini artırmaktadır.

Sonuç olarak, bu çalışma Türkçe metinlerde token seviyesi sorgu-metin eşleştirme için etkili bir hibrit mimari sunmaktadır. Model, verimlilik ve performans dengesi kurarak pratik uygulamalar için ölçeklenebilir bir çözüm sağlamaktadır.

Anahtar Kelimeler: Token Sınıflandırma, Sıfır Örnekle Öğrenme, BERT, Türkçe Doğal Dil İşleme, Semantik Benzerlik, Geç Etkileşimli Füzyon, Focal Loss, Transformer

ABSTRACT

Token-Level Detection of Query-Relevant Sections in Turkish Texts Using Late-Interaction Fusion Model

This thesis addresses the problem of automatically detecting and highlighting query-relevant segments in Turkish texts. The main objective is to develop an effective token-level classification model using zero-shot learning, addressing data scarcity and efficiency issues in current literature.

Unlike traditional information access systems that operate at the document level, this study requires detailed analysis at the token level. This approach aims to determine exactly which words respond to user queries, providing more precise results. For agglutinative languages like Turkish, such analyses present additional challenges due to morphological richness.

Literature review revealed that existing studies focus primarily on English texts with insufficient resources for Turkish. Most existing solutions use either computationally expensive cross-encoder architectures or lower-performing bi-encoder approaches.

The proposed approach utilizes a hybrid architecture called "Late-Interaction Fusion" built upon the `artiwise-ai/modernbert-base-tr-uncased` base model. This architecture consists of two stages: first, queries and texts are encoded independently; second, rich interaction is established through a specialized fusion layer. This approach maintains bi-encoder efficiency while capturing cross-encoder performance.

A Focal Loss-based training strategy addresses class imbalance problems by helping the model focus on challenging examples and balancing positive and negative tokens.

The dataset was created by labeling 500 thousand Turkish texts, compiled from various online sources, using a semi-supervised approach powered by Large Language Models (LLMs) developed in this thesis. This process resulted in approximately 2 million labeled query-text pairs. Data was compiled from news articles, technical blogs, product reviews, and forum posts.

Model training was conducted on 2 NVIDIA H100 GPUs for 10 epochs. After training, the best checkpoint (epoch 2) was selected based on the highest validation F1, using a batch size of 32, a learning rate of 2e-5, and the AdamW optimizer.

To fully analyze the model's learning dynamics, the training process ran for 10 epochs. The validation performance peaked at epoch 2; therefore, all final evaluations

report results from this best checkpoint.

Experimental results show the proposed model, at its best checkpoint (epoch 2), achieved an F1-score of 0.557, significantly outperforming the ‘Keyword Matching‘ (0.065) and ‘Classic Bi-Encoder‘ (0.481) baselines. Furthermore, its performance is highly competitive with the SOTA reference model, ‘GLiNER‘ (0.580).

In terms of efficiency, the proposed model uses fewer parameters and offers a faster inference speed compared to the more complex GLiNER architecture, highlighting its potential for practical applications.

In conclusion, this study presents an effective hybrid architecture for token-level query-text matching in Turkish texts. The model achieves a successful balance between efficiency and performance, providing a scalable solution for practical applications.

Keywords: Token Classification, Zero-Shot Learning, BERT, Turkish Natural Language Processing, Semantic Similarity, Late-Interaction Fusion, Focal Loss, Transformer

KISALTMALAR

API : Application Programming Interface (Uygulama Programlama Arayüzü)

BCE : Binary Cross-Entropy (İkili Çapraz Entropi)

BERT : Bidirectional Encoder Representations from Transformers (Transformerlar'dan Çift Yönlü Kodlayıcı Temsilleri)

BLT : Byte Latent Transformer (Byte Gizli Transformers)

BM25 : Best Matching 25 (En İyi Eşleştirme 25)

ByT5 : Byte-level Text-to-Text Transfer Transformer (Bayt Seviyesi Metinden Metne Aktarım Transformer'ı)

DPR : Dense Passage Retrieval (Yoğun Pasaj Getirme)

FAISS : Facebook AI Similarity Search (Facebook AI Benzerlik Arama)

GLiNER : Generalist Lightweight Model for Various Information Extraction Tasks (Çeşitli Bilgi Çıkarım Görevleri için Genelci Hafif Model)

GPU : Graphics Processing Unit (Grafik İşlem Birimi)

LLM : Large Language Model (Büyük Dil Modeli)

MLP : Multi-Layer Perceptron (Çok Katmanlı Algılayıcı)

NLP : Natural Language Processing (Doğal Dil İşleme)

PII : Personally Identifiable Information (Kişisel Olarak Tanımlanabilir Bilgi)

SOTA : State-of-the-Art (En son bilimsel/teknolojik gelişme)

TARS : Task-Aware Representation of Sentences (Görev Farkındalıklı Cümle Temsili)

UTF-8 : Unicode Transformation Format 8-bit (Unicode Dönüştürme Formatı 8-bit)

XLM-R : Cross-lingual Language Model - RoBERTa (Çapraz Dilli Dil Modeli - RoBERTa)

SEMBOLLER

α	: Öğrenme oranı (learning rate)
β	: Momentum parametresi
θ	: Model parametreleri
σ	: Sigmoid aktivasyon fonksiyonu
$\mathbf{H}$	: Gizli durum matrisi (hidden state matrix)
$\mathbf{W}$	: Ağırlık matrisi (weight matrix)
$\mathbf{b}$	: Bias vektörü
d_{model}	: Model boyutu
n	: Örneklem boyutu
P	: Precision (Kesinlik)
R	: Recall (Duyarlılık)
F_1	: F1-Score

TABLO LİSTESİ

Tablo 3.1. Sıralı Kelime Grubu Eşleştirme Algoritması 34
Tablo 3.2. Anlamsal Arama (Dense Retrieval) Algoritması 35
Tablo 3.3. Geç Etkileşimli Füzyon Modeli Algoritması 36
Tablo 3.4. Model Eğitimi için Kullanılan Hiperparametreler 39

Tablo 4.1. Karşılaştırmalı Model Performans Sonuçları 48
Tablo 4.2. Model Mimarilerinin Karşılaştırmalı Verimlilik Analizi 50

ŞEKİL LİSTESİ

Şekil 3.1. ZSCAT Platformu: Çok-Kaynaklı Anotasyon Arayüzü. İnsan (yeşil), LLM (mor) ve model (mavi) etiketleri aynı anda gösterilmektedir. 29
Şekil 3.2. ZSCAT Platformu: Model Listesi Paneli. Eğitilmiş modeller ve F1 skorları bu arayüzde karşılaştırılmaktadır. 31
Şekil 3.3. ZSCAT Platformu: Model Eğitimi Ayarları. Focal Loss gibi gelişmiş hiperparametreler bu form üzerinden ayarlanmaktadır. . . . 32
Şekil 3.4. ZSCAT Platformu: Canlı Eğitim Durumu Paneli. Eğitim süreci bu arayüzden takip edilmektedir. 33
Şekil 3.5. Önerilen Geç Etkileşimli Füzyon Modeli Mimarisi 36
Şekil 3.6. Önerilen Geç Etkileşimli Füzyon Modeli: Eğitim ve Doğrulama F1 Eğrileri 40

Şekil 4.1. Önerilen Füzyon Modelinin Eğitim Dinamikleri 48
Şekil 4.2. Çift Kodlayıcı Modelinin Eğitim Dinamikleri 48
Şekil 4.3. GLiNER Modelinin Eğitim Dinamikleri 49

1. GİRİŞ

1.1 Problemin Tanımı

Günümüzde kurumsal arama motorlarından müşteri destek sistemlerine, hukuki metin analizinden akademik araştırmalara kadar geniş bir yelpazede, kullanıcılar devasa metin yığınları içinde spesifik bilgilere anında erişme ihtiyacı duymaktadır. Geleneksel anahtar kelime tabanlı arama motorları, bir dokümanın ilgili olup olmadığını belirlemede başarılı olsa da, kullanıcının asıl aradığı cevabın veya kanıtın geçtiği o birkaç cümlelik kritik pasajı bulup doğrudan sunmada yetersiz kalmaktadır. Örneğin, büyük ve sayıca fazla metinlerde belirli bir konuya dair geçen yazıların vurgulanması gerektiğinde, geleneksel anahtar kelime araması yetersiz kalmaktadır. Müşteri geri bildirimleri arasında "teslimat gecikmesi" veya "ürün kalitesi" gibi anlamsal kavramları içeren pasajları tespit etmek, geleneksel anahtar kelime araması ile mümkün değildir çünkü kullanıcılar bu kavramları farklı kelimelerle ifade edebilir. Bu durum, çoğu zaman son derece zaman alıcı ve hataya açık manuel bir sürece dönüşmektedir. Bu tez çalışması, tam olarak bu endüstriyel ve pratik probleme odaklanarak, bir sorgu ile bir metin verildiğinde, metin içerisindeki anlamsal olarak en uyumlu kısımların gerçek zamanlı olarak, token seviyesinde tespit edilip vurgulanmasını sağlayan bir derin öğrenme modeli geliştirmeyi hedeflemektedir.

1.2 Amaç ve Kapsam

Bu çalışmanın temel amacı, BERT tabanlı sinir ağı mimarilerini kullanarak, bir metin ve bir sorgu verildiğinde, metnin hangi kısımlarının (token'larının) sorguyla anlamsal olarak ilişkili olduğunu tespit eden bir "Sıfır Örnekli Token Sınıflandırma" (Zero-Shot Token Classification) modeli geliştirmektir.

Bu kapsamda, modelin eğitimi için herhangi bir göreve özel etiketli örneğe ihtiyaç duymaması (zero-shot öğrenme) hedeflenmektedir. Geliştirilecek olan model, metindeki her bir token için sorguyla uyumlu olup olmadığına dair bir olasılık skoru üretecektir. Bu skorlar kullanılarak, metin içerisindeki ilgili kısımlar belirlenecek ve kullanıcıya sunulmak üzere vurgulanacaktır. Çalışma, aynı temel BERT modeli üzerinde geliştirilen farklı yaklaşımların (Klasik Çift Kodlayıcı, Geç Etkileşimli Füzyon) performansını, literatürdeki temel ve SOTA (state-of-the-art) modellerle karşılaştırmalı olarak analiz ederek en uygun mimariyi belirlemeyi de kapsamaktadır.

1.3 Araştırmanın Önemi ve Literatüre Katkısı

Bu tez çalışması, özellikle Türkçe doğal dil işleme alanında, çeşitli açılardan önemli boşlukları doldurmayı ve literatüre somut katkılar sunmayı hedeflemektedir. Araştırmanın önemi, aşağıdaki temel katkılar etrafında şekillenmektedir:

1. **Hibrit Etkileşim Mimarisi:** Bu tez, literatürdeki verimlilik odaklı çift kodlayıcı ve başarım odaklı çapraz kodlayıcı mimarileri arasında bir denge sunan, geç etkileşimli bir füzyon katmanına sahip hibrit bir yaklaşımın etkinliğini, token-seviyesi alaka tespiti gibi hassas bir görev üzerinde test etmektedir. Bu, verimlilik ve başarım ödünleşiminde (trade-off) yeni bir bakış açısı sunacaktır.

2. **Türkçe İçin Büyük Ölçekli ve Zenginleştirilmiş Veri Seti:** Bu tez çalışması kapsamında, bu spesifik görev için tasarlanmış, Büyük Dil Modeli (LLM) destekli veri zenginleştirme teknikleriyle 2 milyon örneğe ulaştırılmış büyük ölçekli bir Türkçe veri seti oluşturulmuştur. Bu veri seti, hem bu tezin temelini oluşturmakta hem de gelecekteki araştırmalar için Türkçe doğal dil işleme topluluğuna değerli bir kaynak olarak sunulmaktadır.

3. **SOTA Modellerle Karşılaştırmalı Analiz:** Geliştirilen görev-odaklı modelin performansı, GLiNER gibi alanındaki en modern ve genel amaçlı modellere karşı kıyaslanarak, daha hafif ve odaklı mimarilerin karmaşık problemlerdeki rekabetçiliği ve potansiyel avantajları ortaya konulmaktadır.

1.4 Araştırma Soruları ve Hipotezler

Bu çalışmanın temel araştırma soruları şu şekilde formülize edilmiştir:

1. **Ana Araştırma Sorusu:** BERT tabanlı geç etkileşimli füzyon modeli, Türkçe metinlerde token seviyesinde sorgu-uyumluluk tespitinde ne ölçüde başarılı olmaktadır?

2. **Alt Araştırma Sorusu 1:** Önerilen hibrit model, standart çift kule (bi-encoder) mimarisine kıyasla ne kadar performans artışı sağlamaktadır?

3. **Alt Araştırma Sorusu 2:** Az sayıda etiketli veri ile eğitilen model, genel amaçlı büyük ölçekli modellere (GLiNER gibi) kıyasla ne kadar rekabetçi performans göstermektedir?

4. **Alt Araştırma Sorusu 3:** Focal Loss kayıp fonksiyonunun sınıf dengesizliği problemine katkısı ne ölçüdedir?

Temel Hipotez: Sorgu ve metin arasında derin etkileşim kuran geç füzyon mimarisi, çift kodlayıcı yaklaşımlara kıyasla token seviyesinde anlamsal alaka tespitinde anlamlı ölçüde daha başarılı performans sergileyecektir.

1.5 Araştırma Tanımları

Bu çalışmada kullanılan temel kavramların tanımları aşağıda verilmiştir:

Token Sınıflandırma: Bir metnin her bir alt-birimi (token) için, belirli bir kategoriye ait olup olmadığını belirleme görevidir. Bu çalışmada, her token'ın bir sorguyla "ilgili" veya "ilgisiz" olduğu sınıflandırılmaktadır.

Sıfır Örnekli Öğrenme (Zero-Shot Learning): Bir makine öğrenmesi modelinin, eğitim sırasında hiç görmediği sınıfları veya görevleri, sadece o sınıfların/görevlerin bir tanımını alarak tanıyabilmesi yeteneğidir.

Dense Passage Retrieval (Yoğun Parça Getirme): Sorgu ve dokümanları yoğun vektör uzayında temsil ederek, anahtar kelime eşleşmesi yerine anlamsal benzerlik üzerinden arama yapma yöntemidir.

Geç Etkileşimli Füzyon: Sorgu ve metin önce ayrı ayrı kodlandıktan sonra, özel bir dikkat mekanizması ile etkileşime sokulması yaklaşımıdır.

Transformer: Dikkat mekanizması üzerine kurulu, sıralı veri işlemede devrim yaratan sinir ağı mimarisidir.

Bilgi Erişimi (Information Retrieval): Manning vd. (2008) [1] tarafından tanımlandığı üzere, bilgi erişimi, kullanıcıların bilgi ihtiyaçlarını karşılamak amacıyla büyük metin koleksiyonlarından ilgili dokümanları bulma sürecidir. Geleneksel bilgi erişimi sistemleri, doküman düzeyinde alaka tespiti yaparken, bu tez çalışması bu süreci token düzeyine indirgeyerek daha hassas ve detaylı bir alaka analizi sunmayı hedeflemektedir. Bu yaklaşım, Manning'in tanımladığı temel bilgi erişim prensiplerini korurken, kullanıcıların spesifik bilgi ihtiyaçlarını daha etkin bir şekilde karşılamayı mümkün kılmaktadır.

Pratik Uygulama Potansiyeli: Geliştirilen model, kurumsal arama motorları, müşteri hizmetleri analiz platformları, hukuki metin inceleme sistemleri ve akademik literatür tarama araçları gibi birçok pratik uygulamada doğrudan kullanılabilir. Modelin bu pratik değeri, benzer BERT tabanlı yaklaşımların müşteri geri bildirim analizi [2] ve hukuki metin incelemesi [3] gibi endüstriyel alanlarda başarılı bir şekilde kullanılmasıyla da kanıtlanmıştır.

Hibrit Etkileşim Mimarisi: Bu tez, literatürdeki verimlilik odaklı çift kodlayıcı ve başarım odaklı çapraz kodlayıcı mimarileri arasında bir denge kuran "Geç Etkileşimli Füzyon" (Late-Interaction Fusion) yaklaşımını benimser. Bu hibrit yapı, hem yüksek çıkarım hızı potansiyeli sunar hem de sorgu ile metin arasında derin anlamsal bir etkileşim kurarak yüksek başarım elde etmeyi hedefler.

1.6 Sınırlılıklar ve Varsayımlar

Bu araştırmanın kapsamı ve metodolojisi, belirli sınırlılıklar ve varsayımlar çerçevesinde şekillendirilmiştir.

1.6.1 Sınırlılıklar

- **Token Limiti:** Kullanılan Transformer modelinin (BERT) 512 token'lık girdi limitine bağlı olarak, analizler bu limiti aşmayan metin parçaları üzerinde yoğunlaşmıştır.
- **Dil Kısıtı:** Çalışma özellikle Türkçe metinler üzerine odaklanmış olup, diğer dillerdeki genellenebilirlik test edilmemiştir.
- **Alan Kısıtı:** Bu çalışmadaki veri seti, *Materyal ve Yöntem* bölümünün *Veri Seti ve Hazırlık Süreci* alt bölümünde detaylandırıldığı gibi, özellikle müşteri geri bildirimleri gibi endüstriyel değeri yüksek metinlere stratejik olarak odaklanmıştır. Bu nedenle, modelin akademik makaleler veya edebi metinler gibi farklı alanlardaki performansı bu çalışmanın kapsamı dışındadır.

1.6.2 Varsayımlar

- Veri setinin zenginleştirilmesinde kullanılan LLM'in ürettiği sorgu ve etiketlerin, kabul edilebilir düzeyde gürültü içerdiği varsayılmaktadır.
- Internet kaynaklarından toplanan verinin, gerçek kullanım senaryolarını temsil ettiği varsayılmaktadır.
- Etiketleme sürecinde kullanılan LLM'in, insan etiketleyicilerin vereceği kararlara yakın kalitede çıktı ürettiği varsayılmaktadır.

2. KAVRAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR

Bu bölümde, çalışmanın teorik altyapısını oluşturan temel kavramlar ve literatürdeki güncel yaklaşımlar incelenmektedir. Araştırmamız, "az örnekle token sınıflandırma kullanarak metin içerisinde sorguya uygun kısımların tespiti" problemine odaklandığından, bu bölüm özellikle yoğun pasaj getirme (passage retrieval) modelleri ve bu modellerin temelindeki etkileşim mekanizmaları üzerine yoğunlaşacaktır.

Modern yoğun pasaj getirme alanındaki modeller, temel olarak verimlilik ve başarım arasındaki bir ödünleşim (trade-off) ekseninde konumlanmaktadır. Bu eksenin iki zıt kutbunda, literatüre yön veren iki ana mimari bulunur: çift kodlayıcı (Bi-Encoder) ve çapraz kodlayıcı (Cross-Encoder). Bu bölüm, bu iki temel paradigmayı, aralarındaki boşluğu dolduran Geç Etkileşim yaklaşımlarını ve çalışmamızın bu ekosistemdeki yerini detaylandıracaktır.

2.1 Etkileşim Mimarilerinde Temel Paradigmalar

2.1.1 Çift Kodlayıcı (Bi-Encoder) Mimarileri: Verimlilik Odaklı Yaklaşım

Sentence-BERT (SBERT) [4] ile popülerleşen çift kodlayıcı (veya çift kule) modelleri, hesaplama verimliliği ve ölçeklenebilirlik üzerine kurulmuştur. Bu yaklaşımda, sorgu ve dokümanlar, birbirlerinden tamamen habersiz bir şekilde, iki ayrı Transformer kodlayıcıdan geçirilir. Her biri için tek bir yoğun vektör (embedding) üretilir ve bu vektörler arasındaki kosinüs benzerliği gibi bilinen temel bir metrikle alaka skoru hesaplanır.

Bu mimarinin en büyük avantajı, tüm dokümanların vektör temsillerinin önceden hesaplanıp (offline) bir dizinde (index) saklanabilmesidir. Bu sayede, sorgu anında milyonlarca doküman arasından anlamsal arama yapmak milisaniyeler içinde mümkün olur. Ancak bu verimliliğin önemli bir bedeli vardır: etkileşim eksikliği. Sorgu ve doküman tamamen ayrı kulelerde işlendiği için, aralarındaki ince anlamsal ilişkiler (örneğin, kelime sıralamasının veya bir kelimenin diğerinin bağlamındaki anlamının önemi) model tarafından yakalanamaz. Bu durum, başarımın genellikle daha karmaşık modellere göre düşük kalmasına neden olur.

Bu temel kısıtlılık, sorgu ve dokümanı tek bir vektöre sıkıştırma zorunluluğun-

dan kaynaklanır ve literatürde "temsil darboğazı" (representation bottleneck) olarak bilinen bir soruna yol açar [5]. Model, metinler arasındaki zengin ve çok yönlü anlamsal etkileşimi tek bir noktada özetlemek zorunda kaldığı için kaçınılmaz olarak bilgi kaybeder. Bu tezin önerdiği "geç etkileşim" mimarisi, tam olarak bu darboğazı aşmayı hedefler.

2.1.2 Çapraz Kodlayıcı (Cross-Encoder) Mimarileri: Başarım Odaklı Yaklaşım

Spektrumun diğer ucunda, başarımda en üst noktayı hedefleyen çapraz kodlayıcı modelleri yer alır. Bu mimaride, sorgu ve metin en başta birleştirilerek tek bir girdi olarak modele sunulur. Bu sayede, Transformer'ın her katmanında sorgu ve metin token'ları arasında derinlemesine bir çapraz dikkat (cross-attention) etkileşimi gerçekleşir. Bu zengin etkileşim, alaka düzeyini çok daha hassas bir şekilde modellemeyi sağlar ve genellikle en yüksek başarıyı getirir.

Ancak bu üstün başarının hesaplama maliyeti çok yüksektir. Önceden bir dizin oluşturma imkanı yoktur; her bir sorgu-doküman çifti için modelin baştan sona çalıştırılması gerekir. Bu $O(n^2)$ karmaşıklık, çapraz kodlayıcıları büyük ölçekli ve gerçek zamanlı arama senaryoları için pratik olarak kullanılamaz kılar.

2.1.3 Geç Etkileşimli (Late Interaction) Mimariler

Çift kodlayıcıların verimliliği ile çapraz kodlayıcıların başarımı arasındaki bu boşluğu doldurmak amacıyla literatürde Geç Etkileşim (Late Interaction) mimarileri önerilmiştir. Bu alandaki öncü çalışmalardan biri olan ColBERT [6], bu iki yaklaşımı hibrit bir şekilde birleştirerek yeni bir çığır açmıştır. Bu temel mimarinin başarısı, hem modelin kendisinin ColBERTv2 [7] gibi daha verimli versiyonlarının geliştirilmesine, hem de PLAID [8], FIID [9] ve SPLADE++ [10] gibi bu paradigma üzerine inşa edilen ve farklı optimizasyonlar sunan birçok yeni modelin literatüre kazandırılmasına öncülük etmiştir.

ColBERT mimarisinde, sorgu ve dokümanlar önce çift kodlayıcı gibi ayrı ayrı kodlanır. Ancak, her biri için tek bir özet vektör üretmek yerine, her bir token için ayrı bir temsil vektörü (token embeddings) saklanır. Etkileşim ise en sonda, bu iki vektör matrisi arasında, ucuz ama etkili bir 'MaxSim' operatörü ile gerçekleştirilir. Bu "geç" ve "token-bazlı" etkileşim, çapraz kodlayıcıların derinlemesine analiz yeteneğini taklit ederken, çift kodlayıcıların dizinleme ve hızlı arama avantajlarını büyük ölçüde korur.

Bu tezde önerilen model de bu modern geç etkileşim paradigmasını temel almaktadır. Sorgu ve metin için üretilen token-seviyesi temsilleri, 'MultiheadAttention' gibi hafif bir mekanizma ile etkileşime sokarak, Türkçe dilinin yapısına uygun, verimli ve etkili bir token sınıflandırma çözümü sunmayı hedeflemektedir. Bu çalışma, ColBERT ile popülerleşen geç etkileşimli stratejinin, doküman getirme görevlerinden farklı olarak doğrudan token sınıflandırma görevlerindeki potansiyeli ve etkinliği üzerine odaklanmaktadır. Literatürde yeterince incelenmemiş olan bu araştırma boşluğunun doldurulması hedeflenmektedir.

2.2 Transformer Mimarisi ve Dikkat Mekanizması

Günümüzün önde gelen doğal dil işleme modellerinin temelini oluşturan Transformer mimarisi, Vaswani vd. (2017) [11] tarafından "Attention is All You Need" makalesiyle literatüre kazandırılmıştır. Bu devrimci mimari, geleneksel ardışık işleme (sequential processing) yaklaşımlarını tamamen terk ederek, dikkat mekanizması (attention mechanism) tabanlı paralel bir işleme modeli önermiştir.

2.2.1 Transformer'ın Temel Bileşenleri

Transformer mimarisi, çalışmamızın da temelini oluşturan üç ana mekanizmayı literatüre kazandırmıştır:

Öz-Dikkat (Self-Attention) Mekanizması

Transformer'ın en yenilikçi özelliği, her token'ın sekansın tüm diğer token'larıyla doğrudan etkileşim kurabilmesini sağlayan öz-dikkat mekanizmasıdır. Bu mekanizma, üç temel bileşen üzerinden çalışır:

- **Query (Q):** Bir token'ın "neyi aradığını" temsil eden vektör
- **Key (K):** Her token'ın "ne sunduğunu" tanımlayan vektör
- **Value (V):** Gerçek bilgi içeriğini taşıyan vektör

Dikkat skoru şu formülle hesaplanır:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

Bu formüldeki $\sqrt{d_k}$ ölçeklendirme faktörü, boyut arttıkça dikkat skorlarının aşırı büyümesini engelleyerek gradyan stabilitesini sağlar.

Çok Başlıklı (Multi-Head) Dikkat

Tek bir dikkat mekanizması yerine, Transformer paralel olarak çalışan birden fazla "dikkat başlığı" kullanır. Bu çok başlıklı dikkat (İng. *Multi-Head Attention*) mekanizması, her başlığın, girdi sekansının farklı anlamsal yönlerine odaklanabilmesini sağlar. Bu, modelin aynı anda hem sözdizimsel hem de anlamsal ilişkileri yakalayabilmesini sağlar.

Pozisyonel Kodlama (Positional Encoding)

Transformer paralel işleme yaptığından, token'ların sekans içindeki konumsal bilgisini ayrıca enjekte etmesi gerekir. Bu, sinüs ve kosinüs fonksiyonları kullanılarak gerçekleştirilen pozisyonel kodlama (İng. *Positional Encoding*) ile sağlanır ve modelin farklı uzunluktaki sekanslarla genelleme yapabilmesini sağlar.

2.2.2 Transformer'ın Doğal Dil İşlemedeki Etkisi

Transformer mimarisi, doğal dil işleme alanında paradigmatik bir değişim yaratmıştır:

- **Paralelleştirme:** Ardışık modellerin (RNN, LSTM) aksine, tüm pozisyonlar aynı anda işlenebilir
- **Uzun Mesafe Bağımlılıkları:** Her token, sekansın herhangi bir yerindeki diğer token'larla doğrudan etkileşim kurabilir
- **Ölçeklenebilirlik:** Mimarinin basitliği, çok büyük modellerin eğitilmesine olanak tanır

Bu özellikler, BERT, GPT serisi, T5 gibi büyük dil modellerinin geliştirilmesine ve günümüzün NLP başarılarına zemin hazırlamıştır.

2.2.3 Aralık Çıkarımı ve Dikkat Mekanizması

Transformer mimarisinin token seviyesinde sınıflandırma görevlerindeki etkinliği, özellikle aralık çıkarımı (span extraction) görevlerinde kanıtlanmıştır. Seo vd. (2016) [12], BiDAF (Bidirectional Attention Flow) modeli ile SQuAD veri setinde aralık çıkarımı görevinde öncü bir başarı elde etmiştir. Bu model, sorgu ve metin arasında çift yönlü dikkat mekanizması kurarak, metin içerisindeki doğru cevap aralığını (span) tespit etmeyi başarmıştır. BiDAF'in temel yeniliği, sorgudan metne (query-to-context) ve metinden sorguya (context-to-query) olmak üzere iki yönlü dikkat

akışı kurmasıdır. Bu yaklaşım, bizim tezimizde benimsediğimiz geç etkileşimli füzyon mimarisinin temel felsefesiyle uyumludur: sorgu ve metin arasında zengin bir etkileşim kurarak token seviyesinde hassas sınıflandırma yapmak.

2.2.4 Çalışmamızla İlişkisi

Bizim geliştirdiğimiz hibrit etkileşim modeli, Transformer'ın dikkat mekanizmasından doğrudan ilham almaktadır. Özellikle, sorgu ve metin token'ları arasındaki çapraz dikkat işlemi, Transformer'ın temel prensiplerini token-seviyesi alaka tespiti problemine uyarlar. Bu sayede, her metin token'ı sorgu ile derinlemesine etkileşim kurarak hassas alaka skorları üretebilir.

2.3 Büyük Dil Modellerinde Alaka Kavramının Yapısal Analizi

Son yıllarda yapılan çalışmalar, Büyük Dil Modelleri'nin (LLM) sadece metin üretmekle kalmayıp, aynı zamanda karmaşık anlamsal ilişkileri ve alaka düzeyini de içsel olarak nasıl modellediğini anlamaya odaklanmıştır. Bu alandaki araştırmalar, genellikle "mekanistik yorumlanabilirlik" (mechanistic interpretability) olarak adlandırılır ve modellerin "kara kutu" doğasını aydınlatmayı hedefler.

Örneğin, Todd ve arkadaşları (2024) [13], GPT-2 gibi modellerde belirli nöronların, metin içindeki alaka düzeyini takip eden ve öngören özel "alaka tespit devreleri" (relevance-tracking circuits) oluşturduğunu göstermiştir. Bu devreler, bir metin parçasının, daha önce gelen bir konuyla ne kadar ilişkili olduğunu belirleyerek, modelin dikkatini ve sonraki kelime tahminlerini yönlendirir. Benzer şekilde, Gurnee ve Tegmark (2023) [14], LLM'lerin içinde dünya olaylarını, coğrafi konumları ve nesneler arasındaki ilişkileri temsil eden içsel "dünya modelleri" (world models) keşfetmişlerdir.

2.4 Gelişmiş Etkileşim Mimarileri ve Optimizasyon Stratejileri

Çift kodlayıcı mimarisinin etkileşim eksikliğini gidermek amacıyla literatürde çeşitli yaklaşımlar önerilmiştir. Bu yaklaşımlar temel olarak etkileşimi daha erken bir aşamaya çekmeyi, daha verimli hale getirmeyi ve optimizasyon hedeflerini iyileştirmeyi amaçlar.

2.4.1 Erken ve Geç Etkileşim Mekanizmaları

Standart çift kodlayıcı modellerinin aksine, geç etkileşimli modeller ve tamamen birleşik çapraz kodlayıcı modeller, sorgu ve doküman arasındaki anlamsal hizalamayı daha zengin bir şekilde modellemeyi hedefler. Çapraz kodlayıcılar, sorgu ve metni en başından birleştirerek her katmanda derinlemesine bir etkileşim sağlarlar ancak her bir sorgu-doküman çifti için ayrı bir model çağrısı gerektirdiklerinden dolayı yavaştırlar. Bizim bu tezde benimsediğimiz geç etkileşimli mimari (Bölüm 4'te detaylandırılmıştır), bu iki dünyanın arasında bir denge kurar: sorgu ve metni önce ayrı ayrı kodlar (verimlilik), ardından tek bir 'MultiheadAttention' katmanı ile sofistike bir etkileşim kurar (başarım).

2.4.2 Liste Bazlı Optimizasyonlar ve CROSS-JEM

Etkileşim mimarilerindeki bir diğer yenilikçi yön, özellikle birden çok adayın (metnin) tek bir sorgu için sıralandığı durumlarda verimliliği artırmaktır. Standart yaklaşımlar her bir adayı sorguyla ayrı ayrı değerlendirirken, CROSS-JEM [15] gibi yeni nesil modeller, liste bazlı (listwise) bir optimizasyon sunar. CROSS-JEM'in temel fikri, bir sorguya ait tüm aday metinlerdeki *benzersiz* token'ları tek bir havuzda toplayıp, sorgu ile birlikte tek bir Transformer çağrısıyla işlemektir. Bu "token birleştirme" (token union) stratejisi, özellikle adaylar arasında yüksek metin benzerliği olduğunda, aynı token'ların tekrar tekrar işlenmesini önleyerek hesaplama maliyetini önemli ölçüde düşürür. Bu yaklaşım, bizimki gibi geç etkileşimli modellerin pratik uygulamalardaki verimliliğini artırmak için önemli bir potansiyel yol haritası sunmaktadır.

2.4.3 Sıralama Odaklı Kayıp Fonksiyonları

Geleneksel kayıp fonksiyonları, pozitif örneklerin benzerlik skorlarını mutlak olarak artırmaya odaklanır. Bu durum, modelin aşırı öğrenmesine (overfitting) ve eğitim verisindeki gürültüye karşı hassas olmasına neden olabilir. Duan vd. (2025) [16], bu sorunu aşmak için sıralama odaklı (ranking-optimized) bir kayıp fonksiyonu önermektedir.

Bu yaklaşımda amaç, mutlak skorları maksimize etmek yerine, pozitif bir örneğin benzerlik skorunun, tüm negatif örneklerin skorlarından anlamlı bir farkla daha yüksek olmasını sağlamaktır. Bu, temelde bir sıralama problemidir ve modelin daha sağlam ve genelleştirilebilir olmasını sağlar. Bizim çalışmamızda da her bir token için bir olasılık (skor) üretilir ve amaç, ilgili token'ların skorlarının ilgisiz olanlardan daha

yüksek olmasını sağlamaktır. Dolayısıyla, bu sıralama odaklı optimizasyon hedefi, bizim modelimizin eğitim stratejisi için de güçlü bir teorik dayanak sunmaktadır.

2.5 Etkileşim Mimarilerinde Güncel Yaklaşımlar ve Alternatifler

Literatürdeki temel etkileşim paradigmalarını (bi-encoder, cross-encoder) ve bizim benimsediğimiz geç-etkileşimli hibrit yaklaşımı tartıştıktan sonra, bu alandaki en güncel araştırma yönelimlerini incelemek, çalışmamızın konumunu daha net bir şekilde belirlemek için önemlidir. Bu yönelimler, temel olarak "verimlilik" ve "girdi temsilinin doğası" olmak üzere iki ana eksende yoğunlaşmaktadır.

2.5.1 Alternatif Girdi Temsilleri: Token-Free Yaklaşımlar

Bu tezde incelenen ve kullanılan tüm modeller, metni öncelikle "token" adı verilen anlamsal alt-parçalara ayıran bir tokenizasyon katmanına dayanmaktadır. Bu yaklaşım, metni sıkıştırarak ve modele önceden bir miktar anlamsal yapı sunarak verimlilik sağlasa da, kendi içinde önemli sınırlılıklar barındırmaktadır. Bunların en önemlileri; eğitim setinde görülmemiş kelimeleri (out-of-vocabulary) etkin bir şekilde işleyememe, farklı diller arasında (özellikle morfolojik olarak zengin dillerde) adil olmayan ve verimsiz bir tokenizasyon yapma potansiyeli ve yazım hataları gibi yüzey seviyesindeki gürültüye karşı kırılganlıktır.

Bu sorunlara kökten bir çözüm olarak, literatürde token-free veya byte-level olarak adlandırılan ve tokenizasyon adımını tamamen ortadan kaldıran yeni bir mimari sınıfı ortaya çıkmıştır. Bu yaklaşımın öncü modellerinden biri olan ByT5 [17], metni token'lara ayırmak yerine, doğrudan ham UTF-8 byte dizisi olarak işler. Bu felsefenin getirdiği temel avantajlar şunlardır:

- **Evrensellik ve Dayanıklılık:** Yaklaşım, 256 byte'lık evrensel ve sabit bir kelime dağarcığına dayanır. Bu sayede herhangi bir dildeki herhangi bir metni (teknik jargon, yazım hataları, yeni kelimeler dahil) teorik olarak kayıpsız bir şekilde işleyebilir. Bu, modeli dillere ve alanlara karşı daha agnostik ve gürültüye karşı doğal olarak daha dayanıklı hale getirir.

Ancak bu esnekliğin önemli bir bedeli vardır: hesaplama maliyeti. Aynı metin, byte dizisi olarak temsil edildiğinde, token tabanlı temsiline göre çok daha uzun bir sekans oluşturur. Transformer mimarisinin işlem maliyetinin sekans uzunluğunun

karesiyle ($O(n^2)$) orantılı olması, bu durumu önemli bir verimlilik engeli haline getirir. Bu temel ödünleşimi (trade-off) yönetmek, token-free modellerin en önemli araştırma odağı olmuştur. Örneğin, ByT5 bu sorunu, parametrelerini decoder'a kıyasla çok daha derin bir encoder ("ağır encoder") yönünde yeniden dağıtarak ele alırken, CANINE [18] gibi daha yeni mimariler farklı ve yenilikçi bir çözüm sunar. CANINE, maliyetli olan derin Transformer katmanlarına girmeden önce, girdi üzerindeki karakter dizisinin uzunluğunu akıllı bir alt-örnekleme (downsampling) mekanizması ile azaltır. Bu sayede, karakter/byte seviyesinde çalışmanın getirdiği sağlamlığı, token tabanlı modellerin verimliliğine daha yakın bir maliyetle birleştirmeyi hedefler. Daha yeni mimariler olan BLT (Byte Latent Transformer) ve SpaceByte ise, "patching" (komşu byte'ları gruplama) ve bu grupları dinamik olarak oluşturma gibi daha sofistike mekanizmalarla bu verimlilik açığını kapatmaya çalışmaktadır.

Sonuç olarak, bizim çalışmamız kanıtlanmış ve yaygın olan token-tabanlı bir yaklaşımı benimsemiş olsa da, byte-seviyesi modellerin sunduğu bu alternatif dünyanın varlığını ve potansiyelini kabul etmek, mevcut metodolojimizin literatürdeki yerini ve gelecekteki olası gelişim yönlerini anlamak açısından kritik öneme sahiptir. Bu, özellikle çok dilli veya gürültülü metinlerle çalışılacak uygulamalar için önemli bir perspektif sunmaktadır.

2.6 Araştırma Boşluğu ve Tezin Konumlandırılması

Mevcut literatür, yoğun belge getirme ve anlamsal alaka modellemesi konularında önemli ilerlemeler kaydetmiş olsa da, belirli alanlarda boşluklar bulunmaktadır:

1. **Token Seviyesinde Hassasiyet:** Mevcut modeller genellikle doküman veya pasaj seviyesinde tek bir alaka skoru üretir. Metin içindeki hangi kısımların sorguyla daha ilgili olduğunu hassas bir şekilde belirleyen ve vurgulayan yaklaşımlar sınırlıdır.

2. **Sıfır Örnekli Öğrenme:** İncelenen modellerin çoğu, belirli görevler için büyük miktarda etiketli veri ile eğitilmektedir. Bir modelin, eğitim sırasında hiç görmediği görevleri veya sorguları, sadece doğal dil tanımlarını kullanarak çözebilme yeteneği üzerine yapılan çalışmalar, bu tezin odak noktasını oluşturmaktadır.

3. **Türkçe Dil Desteği:** Gelişmiş modellerin büyük bir kısmı İngilizce için geliştirilmiştir. Türkçe'nin morfolojik zenginliği ve yapısal farklılıkları, bu model-

lerin doğrudan uygulanmasını zorlaştırmakta ve dil-spesifik çözümlere ihtiyaç duyulmaktadır.

Bu tez, tam olarak bu araştırma boşluklarını hedeflemektedir. Geliştirdiğimiz model, LLM'lerin içsel çalışma mekanizmalarından ilham alarak, sorgu ile metin arasındaki ilişkiyi token seviyesinde ve derin etkileşimle kurar. Bunu, herhangi bir göreve özel etiketli örneğe ihtiyaç duymadan (zero-shot) ve özel olarak Türkçe için geliştirilmiş bir veri seti üzerinde başararak literatüre özgün bir katkı sunmayı amaçlamaktadır.

2.7 Türkçe Doğal Dil İşleme: Morfolojik Zenginlik ve Çözüm Arayışları

Türkçe, eklemeli bir dil olması nedeniyle doğal dil işleme görevlerinde benzersiz zorluklar sunar [19]. Bu zorluklar, modern dil modellerinin Türkçe'ye uygulanmasında özel yaklaşımlar gerektirir ve bu tezin temel motivasyonunu oluşturur.

2.7.1 Türkçe'nin Morfolojik Özellikleri ve NLP Zorlukları

Türkçe'nin en belirgin özelliği, kelime köklerine eklenen çok sayıda ek ile anlamın oluşturulmasıdır. Bu agglutinatif yapı, aynı kökten türetilen kelimelerin sayısını muazzam ölçüde artırır. Örneğin, "ev" kökünden "evimizdeki" gibi karmaşık kelimeler türetilebilir. Bu durum, geleneksel tokenizasyon yöntemlerini zorlar ve kelime dağarcığı (vocabulary) boyutunu patolojik seviyelere çıkarır.

Türkçe'nin agglutinatif yapısı, doğal dil işleme görevlerinde benzersiz zorluklar yaratır. Bu zorluklar şu şekilde kategorize edilebilir:

- **Morfolojik Karmaşıklık:** Bir kelime kökünden türetilebilen kelime sayısının çok yüksek olması
- **Vowel Harmony:** Türkçe'deki ünlü uyumu kurallarının morfolojik analizi karmaşıklaştırması
- **Case Marking:** İsmin yedi farklı hali (nominatif, akuzatif, datif vb.) ve bunların kombinasyonları
- **Derivational Morphology:** Yapım ekleri ile yeni kelimeler türetme sürecinin karmaşıklığı

Bu zorluklar, standart BERT modellerinin Türkçe metinleri etkin bir şekilde işlemesini engeller ve dil-spesifik çözümler gerektirir.

2.7.2 Türkçe BERT Varyantları: Karşılaştırmalı Analiz

Türkçe için geliştirilmiş BERT modelleri, bu morfolojik zorlukları aşmak için farklı stratejiler benimsemiştir. Literatürde en yaygın kullanılan modeller şunlardır:

dbmdz/bert-base-turkish-uncased

Schweter (2020) [20] tarafından geliştirilen bu model, Türkçe Wikipedia ve haber makaleleri üzerinde eğitilmiştir. Model, standart BERT mimarisini kullanarak Türkçe'nin morfolojik zenginliğini yakalamaya çalışır. Ancak, geleneksel tokenizasyon yöntemlerinin kullanılması nedeniyle, agglutinatif yapının getirdiği kelime çeşitliliği sorunu tam olarak çözülememiştir.

ytu-ce-cosmos/turkish-base-bert-uncased

Yıldız Teknik Üniversitesi tarafından geliştirilen bu model [21], Türkçe'nin morfolojik özelliklerini daha iyi yakalamak için özel bir veri seti üzerinde eğitilmiştir. Model, Türkçe'nin agglutinatif yapısını dikkate alan bir tokenizasyon stratejisi benimser ve morfolojik analiz sonuçlarını modelleme sürecine dahil eder.

Artiwise ModernBERT: Modern Mimari ile Türkçe Optimizasyonu

Bu tezde kullanılan Artiwise ModernBERT [22], Warner vd. (2024) [23] tarafından sunulan ModernBERT mimarisinin Türkçe'ye uyarlanmış versiyonudur. Bu model, standart BERT modellerinin aksine, daha verimli bir mimari ve genişletilmiş bağlam penceresi (8192 token) sunar. Türkçe'nin morfolojik zenginliği göz önünde bulundurulduğunda, bu geniş bağlam kapasitesi, karmaşık kelime yapılarının daha iyi anlaşılmasını sağlar.

2.7.3 Türkçe Token Sınıflandırma ve NER: Ortak Zorluklar ve Çözümler

Token seviyesinde sınıflandırma görevlerinin en yaygın ve önemli uygulaması, Named Entity Recognition (NER) görevidir. NER, her token'ın belirli bir varlık türüne (kişi adı, yer adı, organizasyon vb.) ait olup olmadığını belirleyen bir token classification problemidir. Bu nedenle, Türkçe NER çalışmaları, tezinizin odaklandığı token sınıflandırma görevleriyle doğrudan alakalıdır.

Çelik vd. (2020) [24], Türkçe NER görevlerinde karşılaşılan zorlukları ve çözüm yaklaşımlarını detaylı bir şekilde analiz etmiştir. Bu çalışma, Türkçe'nin agglutinatif

yapısının token seviyesinde sınıflandırma görevlerinde nasıl zorluklar yarattığını ve bu zorlukların üstesinden gelmek için geliştirilen yöntemleri ortaya koyar.

Özellikle, Türkçe'de varlık isimlerinin (entity names) morfolojik olarak değişebilmesi (örneğin, "İstanbul" → "İstanbul'da", "İstanbul'a") NER sistemlerinin performansını önemli ölçüde etkiler. Bu sorun, standart token-tabanlı yaklaşımların yetersiz kalmasına neden olur ve daha sofistike çözümler gerektirir. Bu durum, tezinizin ana konusu olan token seviyesinde alaka tespitinde de benzer zorluklar yaratır.

2.7.4 Türkçe Token Sınıflandırma Görevlerindeki Özel Zorluklar

Bu tezin odaklandığı token seviyesinde sınıflandırma görevleri, Türkçe'de özel zorluklar sunar. Agglutinatif yapı nedeniyle, aynı anlamsal içeriğe sahip kelimeler farklı morfolojik formlarda ortaya çıkabilir. Bu durum, token seviyesinde alaka tespitinde tutarlılık sorunları yaratır.

Örneğin, "ev" kelimesi "evimizdeki" şeklinde uzun bir token olarak karşımıza çıkabilir. Bu durumda, sorgu ile "ev" arasındaki anlamsal ilişki, "evimizdeki" token'ının sadece bir kısmında bulunur. Bu, token seviyesinde sınıflandırma görevlerinde özel yaklaşımlar gerektirir.

2.7.5 Çözüm Yaklaşımları ve Gelecek Yönelimler

Türkçe NLP'deki bu zorlukları aşmak için literatürde çeşitli yaklaşımlar önerilmiştir:

- **Morfolojik Analiz Entegrasyonu:** Tokenizasyon sürecine morfolojik analiz sonuçlarının dahil edilmesi
- **Subword Tokenization:** WordPiece gibi alt-kelime tokenizasyon yöntemlerinin Türkçe'ye özel optimizasyonu
- **Character-Level Modeling:** Karakter seviyesinde modelleme ile morfolojik karmaşıklığın aşılması
- **Hybrid Approaches:** Token ve karakter seviyesi yaklaşımların birleştirilmesi

Bu tezde benimsenen "Geç Etkileşimli Füzyon" mimarisi, bu yaklaşımlardan hibrit bir strateji benimser. Model, hem token seviyesinde anlamsal temsilleri yakalar hem de karakter seviyesindeki morfolojik detayları dikkate alır.

2.8 Modern Dil Modellerinin Temel Taşları

Günümüzün büyük dil modelleri, özellikle de bu çalışmanın merkezinde yer alan Transformer mimarisi, kendisinden önce gelen ve alanın temellerini atan devrimci fikirlerin üzerinde yükselmektedir. Transformer'ın paralel işleme ve dikkat mekanizması gibi yeniliklerinin tam olarak anlaşılabilmesi için, bu mimarinin çözdüğü temel problemlerin ve üzerine inşa edildiği kavramların incelenmesi kritik öneme sahiptir.

2.8.1 Ardışık Veri Modelleme ve Uzun Vadeli Bağımlılık Sorunu

Transformer öncesi dönemde, metin gibi ardışık verilerin modellenmesinde Tekrarlayan Sinir Ağları (RNN'ler) ve onların gelişmiş bir versiyonu olan Uzun Kısa Süreli Bellek (LSTM) ağları standart yaklaşım olarak kabul edilmekteydi. LSTM mimarisi, özellikle Hochreiter & Schmidhuber (1997) [25] tarafından sunulan öncü çalışma ile, metin içindeki uzun mesafeli anlamsal bağımlılıkları yakalama konusunda önemli bir adım atmıştır. Ancak bu modellerin en temel kısıtlılığı, veriyi ardışık olarak, yani kelime kelime işlemek zorunda olmalarıydı. Bu durum, hem donanım hızlandırmadan (GPU gibi) tam olarak faydalanmayı engelliyor hem de çok uzun metinlerde "unutma" problemine (vanishing gradient) yol açıyordu. Transformer mimarisi, bu ardışık işleme zorunluluğunu ortadan kaldırarak bu iki temel soruna radikal bir çözüm getirmiştir.

2.8.2 Kelime Temsilleri: Word2Vec ve GloVe

Modern dil modellerinin bir diğer temel taşı, kelimelerin anlamsal özelliklerini yoğun vektörler (dense vectors) olarak temsil etme fikridir. Bu alanda bir çığır açan çalışmalardan biri, Mikolov vd. (2013) [26] tarafından sunulan Word2Vec modelidir. Word2Vec, kelimelerin birlikte kullanım sıklıklarına bakarak, "kral - erkek + kadın" işleminin sonucunun "kraliçe" vektörüne yakın olduğu gibi etkileyici anlamsal ilişkileri vektör uzayında yakalayabilen temsiller öğrenmiştir. Bu fikri bir adım öteye taşıyan Pennington vd. (2014) [27] ise GloVe modelini sunarak, yerel bağlam pencereleri yerine global kelime-kelime birlikte geçme istatistiklerini kullanarak daha sağlam vektörler üretmiştir. BERT gibi bağlamsal (contextual) modeller, bu temel kelime temsil fikirlerini, bir kelimenin anlamının içinde bulunduğu cümleye göre değişebileceği dinamik bir yaklaşımla bir sonraki seviyeye taşımıştır.

2.9 Sıfır Örnekli Öğrenme Paradigmaları ve Varlık İlişkilendirme

Pratik uygulamalarda, özellikle de bizimki gibi dar ve özelleşmiş alanlarda, her olası etiket için binlerce örnek toplamak neredeyse imkansızdır. Eğitim verisinde hiç yer almayan (sıfır örnekli, zero-shot) görevlerle başa çıkabilme yeteneği, modern makine öğrenmesi sistemleri için kritik bir öneme sahiptir. Bu durum, literatürde [28] tarafından "X-Shot Öğrenme" olarak kavramsallaştırılmıştır ve hem az örnekli (few-shot) hem de sıfır örnekli (zero-shot) senaryoları kapsar. Bu tez, bu spektrumun sıfır örnekli ucuna odaklanmaktadır.

2.9.1 Çoklu Görev Tabanlı Eğitim ve Sıfır Örnekli Genelleme

Sıfır örnekli öğrenme alanındaki en önemli gelişmelerden biri, Sanh vd. (2021) [29] tarafından sunulan T0 modelidir. Bu model, çoklu görev üzerinde prompt-tabanlı eğitim (multitask prompted training) ile sıfır örnekli görev genellemesi (zero-shot task generalization) başarısını elde etmiştir. T0'un temel yaklaşımı, farklı NLP görevlerini doğal dil promptlarına dönüştürerek tek bir model üzerinde eğitim yapmaktır. Bu sayede model, eğitim sırasında hiç görmediği yeni görevleri, sadece görev tanımını okuyarak çözebilme yeteneği kazanır. Bu paradigm, bizim tezimizdeki "sıfır örnekli token sınıflandırma" yaklaşımımız için güçlü bir teorik temel sunmaktadır. Özellikle, görev tanımlarının doğal dil formatında verilmesi ve modelin bu tanımları kullanarak genelleme yapması fikri, bizim sorgu-metin etkileşimi modelimizin temel felsefesiyle örtüşmektedir.

Bu zorluğun üstesinden gelmek için, özellikle varlık ilişkilendirme (entity linking) alanında, yani bir metin parçasını bir bilgi tabanındaki (knowledge base) belirli bir varlıkla eşleştirme görevinde, çeşitli paradigmalar geliştirilmiştir.

2.9.2 Çift Aşamalı Yaklaşımlar: Bi-encoder ve Cross-encoder

Varlık ilişkilendirme için standart yaklaşımlardan biri **BLINK** [30] tarafından popüler hale getirilen iki aşamalı modeldir. Bu model, bizim daha önce tartıştığımız çift kule mimarisine benzer bir mantık kullanır:

1. **Aday Getirme (Candidate Retrieval):** Hızlı ama daha az hassas bir **bi-encoder** modeli, büyük bir varlık havuzundan (örneğin Wikipedia'daki tüm

makaleler) sorguyla ilgili olabilecek en iyi K adet adayı (örneğin 64 aday) getirir.

2. **Yeniden Sıralama (Re-ranking):** Yavaş ama çok daha hassas bir **cross-encoder** modeli, bu K adayı tek tek sorguyla birlikte işleyerek en doğru olanı bulur.

Bu mimarinin en büyük avantajı, metin ve varlık temsillerinin ayrı ayrı, önceden hesaplanabilmesidir. Bu verimlilik odaklı yaklaşım, Transformer mimarilerini ölçeklenebilir kılmak için literatürde önerilen nicemleme (quantization) ve budama (pruning) gibi birçok tekniği kapsayan daha geniş bir araştırma alanının parçasıdır [31].

2.9.3 Üretim Tabanlı Yaklaşımlar: Üretimle Getirme

Bu sorunlara bir çözüm olarak, **GENRE** (Generative Entity Retrieval) [32] modeli, problemi bir "üretim" (generation) problemine dönüştürür. Model, metni okur ve ilgili varlığın adını veya ID'sini otoregresif bir şekilde, yani kelime kelime üretir. Bu yaklaşımın getirdiği devrimsel avantajlar şunlardır:

- **Hafıza Verimliliği:** Milyonlarca varlık için ayrı ayrı vektörler saklamak yerine, tüm bilgi modelin parametrelerinde saklanır.
- **Dinamik Varlık Kümesi:** Bilgi tabanına yeni bir varlık eklendiğinde, sadece isminin modele tanıtılması yeterlidir; büyük bir vektör dizinini yeniden oluşturmaya gerek kalmaz.

GENRE, varlıkları önceden tanımlanmış bir liste arasından "seçmek" yerine, doğru varlığı "yazarak" bulur. Bu, bizim çalışmamızdaki "token'ları sınıflandırmak" yerine "ilgili token'ları işaretlemek" fikriyle felsefi bir benzerlik taşır.

2.9.4 Sıfır Örnekli Sınıflandırmanın Yeniden Formüle Edilmesi

Varlık ilişkilendirmenin ötesinde, genel metin sınıflandırma görevleri için de yenilikçi sıfır örnekli yaklaşımlar mevcuttur. TARS (Task-Aware Representation of Sentences) [33], bu alanda farklı bir bakış açısı sunar. Geleneksel modellerin aksine, TARS herhangi bir metin sınıflandırma görevini ikili bir "evet/hayır" sorusuna dönüştürür.

Örneğin, bir metnin "spor" kategorisine ait olup olmadığını anlamak için, model metni ve "Bu metin sporla mı ilgili?" sorusunu birlikte işler ve çıktısı "evet" veya "hayır" olur. Bu sayede model, etiketlerin (kategorilerin) anlamsal içeriğini doğrudan

öğrenir. Bu yaklaşım, yeni bir kategori eklendiğinde (örneğin "siyaset"), modelin yeniden eğitilmesine gerek kalmadan, sadece yeni bir soru ("Bu metin siyasetle mi ilgili?") sorarak çalışmasına olanak tanır. Bu esneklik, az ve sıfır örnekli öğrenme senaryoları için TARS'ı oldukça güçlü kılmaktadır.

2.9.5 Zero-Shot Token Etiketleme Paradigmaları

Cümle veya doküman seviyesinde sınıflandırma yapmak yerine, bu tezin odaklandığı gibi token seviyesinde etiketleme yapmak, farklı ve daha incelikli yaklaşımlar gerektirir. Literatür bu sorunu çözmek için çeşitli paradigmalar önermiştir.

Bu alandaki temel fikirlerden biri, sadece cümle seviyesinde etiketlenmiş verilerle eğitilmiş bir modelin, dikkat mekanizması aracılığıyla hangi token'ların önemli olduğuna dair öğrendiği bilgiyi token seviyesinde etiketleme yapmak için kullanmasıdır. Bujel vd. (2021) [34], bu "dikkat tabanlı aktarım" fikrini sistematik olarak inceleyen ve Transformer tabanlı cümle sınıflandırıcılarının herhangi bir token seviyesi etiket görmeden nasıl etkili birer ardışık etiketleyiciye dönüştürülebileceğini gösteren ilk çalışmalardan biridir.

Bir diğer etkili paradigma, etiketlerin anlamsal içeriğini modelin bir parçası haline getirmektir. Aly vd. (2021) [35], sınıf (etiket) açıklamalarını doğrudan modelin girdisine dahil etmiş ve bir cross-attention mekanizması ile metindeki token'lar ve etiket açıklamaları arasında anlamsal bir ilişki kurarak sıfır örnekli NER görevini başarmıştır. Bu yaklaşım, modelin daha önce hiç görmediği bir varlık türünü, sadece o türün tanımını okuyarak tanıyabilmesini sağlar.

Bu "prompt" (yem) tabanlı fikir, PromptNER [36] gibi daha gelişmiş modellerle bir adım öteye taşınmıştır. PromptNER, hem varlıkların metin içindeki yerini bulma (locating) hem de türünü belirleme (typing) görevlerini, çift yuvalı (dual-slot) bir prompt şablonu içinde birleştirerek tek bir adımda çözer. Bu, çıkarım verimliliğini artırırken, modelin farklı görevler arasında bilgi transferi yapmasını kolaylaştırır.

Veri kıtlığı problemini ele alan tamamlayıcı bir strateji ise kendi kendine öğrenmedir (self-training). Bu paradigmada, bir modelin etiketlenmemiş veri üzerindeki kendi yüksek güvenli tahminleri, "sözde etiket" (pseudo-label) olarak kullanılır ve model bu yeni oluşturulan veri ile tekrar eğitilir. Gera vd. (2022) [37], bu basit ama etkili tekniğin, sıfır örnekli sınıflandırıcıların performansını önemli ölçüde artırdığını göstermiştir.

Bu tezde sunulan model, bu literatürdeki birkaç temel fikirden ilham almaktadır. Özellikle, etiket açıklamalarını kullanarak anlamsal esneklik sağlama (Aly vd. gibi) ve sorgu ile metin arasında zengin bir etkileşim kurarak hassas token-seviyesi puanlama yapma (Bujel vd. ve PromptNER'ın altında yatan mekanizmalar gibi)

yaklaşımlarını birleştiren hibrit bir strateji izlemektedir.

2.9.6 Üretken ve Ayırt Edici Yaklaşımlar

Token seviyesinde sınıflandırma görevlerinde, ayırt edici (discriminative) yaklaşımların yanı sıra üretken (generative) modellerin de etkinliği literatürde kanıtlanmıştır. Raffel vd. (2020) [38], T5 (Text-to-Text Transfer Transformer) modeli ile tüm NLP görevlerini "metin-çıktı" formatına dönüştürerek tek bir üniform çerçevede çözmeyi başarmıştır. T5'in temel felsefesi, her görevi bir çeviri problemi olarak ele almak ve "girdi metni → çıktı metni" şeklinde modellemektir. Bu yaklaşım, token sınıflandırma görevlerinde de uygulanabilir: örneğin, bir metin ve sorgu verildiğinde, model ilgili token'ları doğrudan üretebilir. Bu üretken paradigma, bizim benimsediğimiz ayırt edici yaklaşımdan farklı olarak, token'ları "seçmek" yerine "yazmak" stratejisini benimser. Her iki yaklaşımın da kendi avantajları bulunmakta olup, bizim çalışmamız ayırt edici yaklaşımın verimlilik ve hassasiyet avantajlarından faydalanmayı hedeflemektedir. Bu ayırt edici yaklaşımların yanı sıra, literatürde T5 gibi üretken modelleri bir soru-cevap formatına adapte ederek zero-shot token sınıflandırma görevini çözen alternatif paradigmalar da bulunmaktadır [39].

Bu paradigmalar, ZShot [40] gibi açık kaynaklı çatılar aracılığıyla biyomedikal [41] gibi özel alanlarda dahi etkinliğini kanıtlamıştır. LLM'lerin alaka kavramını nasıl anladığına dair mekanistik çalışmalar [42] devam ederken, belirli klinik senaryolarda optimize edilmiş BERT tabanlı modellerin sıfır örnekli yöntemlere üstün gelebildiği de raporlanmıştır [43], bu da göreve özel ince ayarın önemini koruduğunu göstermektedir.

2.9.7 Çapraz Dilli Modellerde Açık Token Hizalama

Daha önce tartışılan modeller genellikle tek bir dil (İngilizce) odağında geliştirilmiş veya çok dilliliği örtük olarak ele almıştır. Ancak, bir modelin farklı diller arasında gerçekten sağlam temsiller öğrenebilmesi için daha açık hizalama sinyallerine ihtiyaç duyulabilir. Chi vd. (2021) [44], bu yönde önemli bir adım atarak, XLM-R gibi güçlü çok dilli modellerin performansını, yeni bir ön-eğitim görevi ile daha da ileri taşımayı önermiştir.

Bu yaklaşımın merkezinde, gürültü azaltıcı kelime hizalama (denoising word alignment) görevi bulunmaktadır. Model, iki dildeki paralel cümleleri (örneğin, "Kırmızı araba" ve "The red car") alır ve hangi kelimenin hangi kelimeye karşılık geldiğini (araba $\longleftrightarrow$ car) kendi kendine etiketler. Ardından, cümlenin bir kısmı maskelenir ve modelden, bu maskelenmiş girdiye rağmen doğru hizalamayı tahmin etmesi istenir.

Bu süreç, modelin diller arası token seviyesinde çok daha hassas ve gürültüye karşı dayanıklı bir eşleştirme yeteneği kazanmasını sağlar. Bu çalışma, bizim tezimizin temelindeki "token seviyesinde alaka tespiti" fikri için güçlü bir kanıt sunmaktadır: Bir model, token'lar arasındaki anlamsal hizalamayı açıkça modellemek üzere eğitildiğinde, özellikle token seviyesinde muhakeme gerektiren görevlerde (soru-cevap, yapısal tahmin vb.) daha başarılı olmaktadır. Bu, bizim sorgu ve metin token'ları arasında kurmaya çalıştığımız derin etkileşimin, literatürdeki en ileri çok dilli modellerin gelişim yönüyle ne kadar uyumlu olduğunu göstermektedir.

2.9.8 Yoğun Belge Erişimi

Yoğun Belge Erişimi (İng. *Dense Retrieval*), sorgu ve dokümanları yoğun vektör temsillerine dönüştürerek anlamsal olarak eşleştirmeye dayanan modern bir bilgi erişimi yöntemidir. Bu alanda, özellikle DPR (Dense Passage Retriever) [45] gibi temel çalışmalar, çift kuleli (dual-encoder) mimarilerin etkinliğini kanıtlamıştır. Bu temel üzerine inşa edilen RocketQA [46] ve denetimsiz öğrenme yaklaşımları sunan Contriever [47] gibi modeller, bu alandaki araştırma ve geliştirmenin ne kadar dinamik olduğunu göstermektedir. Bu yöntem, anahtar kelime eşleşmesine dayalı geleneksel sparse retrieval yöntemlerine (örn. BM25, TF-IDF) kıyasla, sorgunun anlamsal niyetini daha iyi anlama yeteneği sayesinde üstün sonuçlar sunar.

2.9.9 Passage-Level Evidence ve Token-Level Evidence

Yoğun belge erişimi alanının temel taşlarından biri, Callan (1994) [48] tarafından sunulan bölüm düzeyinde kanıt (passage-level evidence) kavramıdır. Bu çalışma, uzun dokümanlarda arama yaparken, dokümanın tamamı yerine ilgili bölümlerin (İng. passage) kullanılmasının performansı artırdığını göstermiştir. Callan'ın yaklaşımı, dokümanları daha küçük, anlamsal olarak tutarlı bölümlere ayırarak, her bölüm için ayrı ayrı alaka skorları hesaplamayı önerir. Bu fikir, bizim tezimizdeki token düzeyinde kanıt (token-level evidence) yaklaşımımızın teorik temelini oluşturmaktadır. Eğer bölüm düzeyinde kanıt doküman erişiminde başarılıysa, token düzeyinde kanıt da daha hassas ve detaylı bir alaka tespiti için etkili olmalıdır. Bizim modelimiz, Callan'ın bölüm düzeyinde yaklaşımını bir adım öteye taşıyarak, metin içerisindeki her bir token'ın sorguyla olan ilişkisini ayrı ayrı değerlendirmektedir.

2.9.10 Büyük Dil Modelleri ile Veri Zenginleştirme

Büyük Dil Modelleri (LLM) kullanılarak sentetik veri üretimi, özellikle veri kıtlığı yaşanan alanlarda model performansını artırmak için literatürde yaygın olarak kabul

görmüş güçlü bir tekniktir [49, 50].

3. MATERYAL VE YÖNTEM

Bu bölümde, tez kapsamında önerilen modelin mimarisi, bu modelin pratik uygulaması için gereken veri hazırlık süreçleri, modelin eğitim detayları ve performansının ölçüldüğü metrikler detaylı bir şekilde açıklanmaktadır.

3.1 Veri Seti ve Hazırlık Süreci

Bu çalışmanın en önemli yönlerinden biri, Türkçe'ye özel, büyük ölçekli ve hibrit bir yaklaşımla etiketlenmiş bir veri seti üzerinde çalışmasıdır.

3.1.1 Veri Kaynağı ve LLM Destekli Zenginleştirme

LLM, özellikle etiketli veri bulmanın zor ve maliyetli olduğu durumlarda, veri zenginleştirme (data augmentation) için güçlü bir araç olarak ortaya çıkmıştır. Literatürde LLM'lerin bu amaçla kullanımı genellikle iki ana yaklaşıma ayrılmaktadır: (a) tamamen yeni, sentetik metin örnekleri üretmek veya (b) mevcut, gerçek dünya verileri üzerinde LLM'leri bir "veri etiketleyici" (data annotator) veya "akıl yürütücü" (reasoner) olarak kullanmak [49, 50].

Tamamen sentetik veri üretme yaklaşımı, ölçeklenebilirlik açısından cazip olsa da, önemli riskler barındırmaktadır. LLM'ler, ne kadar gelişmiş olurlarsa olsunlar, sıfırdan metin üretirken stilistik olarak tekdüze, belirli kalıpları tekrar eden ve gerçek dünya verisinin karmaşık çeşitliliğini tam olarak yansıtamayan çıktılar verme eğilimindedir [51]. Bu durum, "model çöküşü" (model collapse) gibi, modelin kendi ürettiği verilerle eğitildikçe performansının düşmesi riskini de beraberinde getirir.

Bu potensiye tuzaklardan kaçınmak ve üretilen veri setinin kalitesini ve güvenilirliğini en üst düzeye çıkarmak için, bu tezde bilinçli bir stratejiyle ikinci yaklaşım benimsenmiştir: LLM'i bir "içerik üretici" olarak değil, büyük ölçekli ve yüksek hassasiyetli bir "sıfır örnekli veri etiketleyici" (zero-shot data annotator) olarak konumlandırmak. Bu paradigmanın etkinliği ve LLM'lerin bu görevdeki potansiyeli, literatürdeki öncü ampirik çalışmalarda da gösterilmiştir [52]. Ancak bu tez, basit bir etiketlemenin ötesine geçerek, LLM'leri iki aşamalı özgün bir örnekleme stratejisinin merkezine yerleştirir:

1. **Sıfır Örnekli Konu Çıkarma ve Kanıt Toplama (Pozitif Örnekleme):** Bu aşamada, LLM'den verilen bir metin içerisindeki tüm farklı konuları, herhangi bir önceden tanımlanmış kategori listesi olmaksızın, "sıfır örnekli" bir

yaklaşımla kendisinin tespit etmesi istenmiştir. Bu yöntemin en kritik yönü, modelin tespit ettiği her bir konu için, o konunun metnin hangi bölümünde geçtiğini gösteren kanıtları, metinden birebir alıntılayarak (‘verbatim‘) sunma zorunluluğudur. Bu "kanıt temelli etiketleme" yaklaşımı, hem üretilen konuların metne dayandırılmasını sağlayarak veri bütünlüğünü güvence altına alır hem de etiketleme sürecini şeffaf ve doğrulanabilir kılar.

2. **Dolaylı Negatif Örnekleme:** Modelin, bir metnin ne "olmadığını" da öğrenmesi kritik olduğundan, yüksek kaliteli negatif etiketler üretmek için dolaylı bir yaklaşım benimsenmiştir. Bu aşamada, sisteme entegre bir ‘SmartCandidateManager‘ modülü tarafından seçilen ve metinle alakalı olma olasılığı düşük olan "aday kategoriler" LLM’e sunulmuştur. Modelden, bu aday listesinden hangilerinin metinde "mevcut olduğunu" belirtmesi istenmiştir. Daha sonra, uygulama mantığı, LLM’in "mevcut" olarak işaretlemediği kategorileri, teyit edilmiş "negatif etiket" olarak kabul etmiştir. Bu dolaylı yaklaşım, LLM’in bir kavramın yokluğunu kesin olarak teyit etmesinden ziyade, varlığını teyit etmesinin daha güvenilir bir bilişsel görev olduğu varsayımına dayanmaktadır.

Veri setinin temelini, internet üzerindeki herkese açık kaynaklardan (ürün yorumları, haber makaleleri, forumlar vb.) derlenen yaklaşık 500 bin adet Türkçe metin oluşturmaktadır. Bu metinlerden yüksek kaliteli sorgu-metin çiftleri üretmek amacıyla, LLM-destekli, yarı denetimli bir etiketleme stratejisi izlenmiştir. Sürecin ilk adımında, veri setinin yaklaşık %1’lik bir kısmı uzman insan etiketleyiciler tarafından manuel olarak işaretlenerek bir "altın standart" (gold standard) veri seti oluşturulmuştur. Bu referans seti, etiketleme görevindeki performanslarını ölçmek amacıyla farklı LLM’lerin (örn: Gemma, GPT-4o, GPT-4.1, Qwen3-32b) karşılaştırıldığı bir ön çalışma için kullanılmıştır. Bu karşılaştırma sonucunda, Qwen3-32b modelinin, Türkçe metinler üzerindeki muhakeme (reasoning) yeteneği ve ürettiği etiketlerin insan etiketleriyle olan tutarlılığı bakımından %92’lik bir doğruluk oranına ulaşarak diğer modellere kıyasla en iyi performansı sergilediği tespit edilmiştir. Bu ampirik sonuca dayanarak, kalan yaklaşık 495 bin metnin tamamı, bu tezde geliştirilen iki aşamalı "pozitif ve dolaylı negatif örnekleme" yöntemiyle Qwen3-32b modeli kullanılarak otomatik olarak etiketlenmiştir. Bu süreç sonucunda, başlangıçtaki 500 binlik metin havuzu, yaklaşık 2 milyon etiketli örneğe sahip nihai eğitim veri setine dönüştürülmüştür.

Veri Etiketlemede Kullanılan Prompt Yapıları

Veri etiketleme sürecinin tekrarlanabilirliğini ve şeffaflığını sağlamak amacıyla, kullanılan temel prompt yapıları aşağıda sunulmuştur.

Pozitif Etiketleme Prompt'u Bu prompt, "sıfır örnekli" (zero-shot) bir konu çıkarma ve kanıt toplama görevi gerçekleştirmek üzere tasarlanmıştır.

```
Analyze the following customer review and generate a
   response in JSON format
according to the specified rules.

RULES:
1.  Topic Extraction: Identify ALL DISTINCT TOPICS mentioned
   in the review.
    A review may contain multiple topics; extract as many as
       you can individually.
    The topic titles must be short, concise, and entirely in
       English.
2.  Segment Addition: For each topic, add the sentences or
   phrases where that
    topic is mentioned to the 'segments' list by COPYING
       THEM VERBATIM from the
    review text.
    - VERY IMPORTANT: Segments can NEVER be a summary, a
       category name, or
      your own interpretation. They must be copied verbatim,
         exactly as they
      appear in the text. Failure to follow this rule will
         render the output
      completely invalid.

Required JSON Format:
{{"categories":[{{"category":"<topic_name>","segments":["<copied_segment_f
"<segment_2>"]}}]}}

Customer Review:
{content}
```

Negatif Örnekleme Prompt'u Bu prompt, bir negatif örnekleme görevini gerçekleştirmek üzere tasarlanmıştır. Amacı, önceden seçilmiş bir aday listesinden hangi

kategorilerin verilen metinle ilgili olmadığını tespit etmektir.

```
From the list of categories below, WHICH ONES are ACTUALLY
   present in this
customer review?
List: {', '.join(negative_candidates)}

Return ONLY the following JSON:
{{"present": ["<present_category1>", "<present_category2>"]}}

Customer review: {content}
```

3.1.2 Evren ve Örneklem: Endüstriyel Odaklı Veri Seti

Bu çalışmanın evrenini, modern Doğal Dil İşleme (DDİ) uygulamaları için hem büyük bir zorluk hem de yüksek ticari değer taşıyan, yapısal olmayan, gürültülü ve yüksek hacimli kullanıcı metinleri oluşturmaktadır. Bu doğrultuda örneklem, genel amaçlı internet metinlerinden (haber, blog vb.) bilinçli bir şekilde kaçınılarak, tezdeki problemin doğasına en uygun veri türü olan müşteri geri bildirimlerine stratejik olarak odaklanılarak derlenmiştir.

Veri setinin, tek bir endüstriye veya ürüne bağlı kalarak dar bir alanda uzmanlaşmasını önlemek amacıyla, e-ticaret, telekomünikasyon, finans ve hizmet sektörü gibi çok çeşitli alanlardan toplanan müşteri destek talepleri, ürün/hizmet yorumları ve şikayet metinlerini kapsaması sağlanmıştır. Bu veri türünün seçilmesinin temel nedeni, kullanıcıların bir sorunu veya ihtiyacı (örtük bir sorgu) doğal dilde ifade ettiği ve metnin geri kalanının bu sorguya dair kanıtlar veya detaylar sunduğu, tezdeki "anlamsal token eşleştirme" görevine mükemmel bir şekilde uyan doğal bir laboratuvar olmasıdır. Bu odaklanmış yaklaşım, geliştirilen modelin teorik bir konsept olmaktan çıkıp, endüstriyel değeri yüksek, gerçek dünya verisinin karmaşıklığıyla başa çıkabilen bir çözüm olarak konumlandırılmasını sağlamıştır.

Nihai veri seti, bu kaynaklardan derlenen metinler üzerinde LLM-destekli etiketleme sonucunda oluşan 2 milyon sorgu-metin çiftini içermekte olup, bu çiftlerin token seviyesinde etiketlenmesiyle toplamda yaklaşık 50 milyon token düzeyinde etiket elde edilmiştir. Her bir metin parçası maksimum 512 token uzunluğunda sınırlandırılmıştır.

3.1.3 Verilerin Toplanması

Veri toplama süreci 2024 yılının ilk çeyreğinde gerçekleştirilmiştir. Ham veriler, herkese açık API'ler ve web scraping teknikleri kullanılarak toplandıktan sonra, kişisel bilgi içeren veriler filtrelenmiştir.

LLM destekli etiketleme süreci, kontrollü bir ortamda yürütülmüş ve her bir otomatik etiketin kalite kontrol sürecinden geçmesi sağlanmıştır. Bu süreçte, rastgele seçilen 1000 örnek üzerinde manuel kontrol yapılmış ve otomatik etiketleme doğruluk oranının %92 düzeyinde olduğu tespit edilmiştir.

3.1.4 Veri Ön İşleme ve Tokenizasyon

Modelin eğitimi için veri hazırlığı kritik adımlar içerir:

1. **Tokenizasyon:** Tüm sorgu ve metinler, temel aldığımız `artiwise-ai/modernbert-base-tr-uncased` modelinin kendi WordPiece tokenizer'ı ile alt-parçalara (subword) ayrılmıştır. Bu yaklaşım, Türkçe'nin eklemeli yapısından kaynaklanan kelime çeşitliliğini etkin bir şekilde yönetir.

2. **Uzun Metin Yönetimi:** BERT mimarisinin 512 token'lık girdi limitini aşan metinler ve sorgular, basitçe ilk 512 token'ı alınarak kısaltılmıştır.

3.1.5 Temel Dil Modeli Seçimi: Artiwise ModernBERT

Bu çalışmanın temelini oluşturan dil modelinin seçimi, projenin başarısı için kritik bir adımdır. Piyasada birçok Türkçe BERT tabanlı model bulunmasına rağmen, bu tezde temel model olarak **Artiwise ModernBERT** [22] tercih edilmiştir. Bu model, Warner vd. (2024) tarafından sunulan ve BERT mimarisini modern optimizasyonlarla bir sonraki seviyeye taşıyan **ModernBERT** [23] temel alınarak geliştirilmiştir. Bu stratejik seçimin arkasında yatan temel nedenler, hem ModernBERT'in arkasındaki akademik yeniliklere hem de ArtiwiseAI tarafından geliştirilen Türkçe versiyonunun kanıtlanmış performansına dayanmaktadır:

- **Modernleştirilmiş Mimari ve Verimlilik:** [23] makalesinde belirtildiği gibi, ModernBERT, standart BERT mimarisine kıyasla hem hız hem de bellek kullanımı açısından önemli iyileştirmeler sunar. Bu, özellikle büyük veri setleri üzerinde yapılan eğitim (fine-tuning) ve çıkarım süreçlerini daha verimli hale getirir.

- **Genişletilmiş Bağlam Penceresi:** Geleneksel BERT modelleri 512 token'lık bir bağlam penceresi ile sınırlıyken, ModernBERT mimarisi bu limiti doğal

olarak 8192 token'a çıkarır [23]. Bu çalışmada standart 512 token limiti kullanılmış olsa da, modelin bu doğal uzun bağlam kapasitesi, daha karmaşık ve uzun metinlerdeki anlamsal ilişkileri yakalama konusunda temel bir potansiyel sunmaktadır.

- **Türkçe için Kanıtlanmış Üstün Performans:** ArtiwiseAI tarafından yayınlanan ve bizim de temel aldığımız modelin Hugging Face üzerinde sunulan benchmark sonuçları, modelin Türkçe için etkinliğini kanıtlamaktadır [22]. Model, çeşitli Türkçe anlama görevlerinde, [21] ve [20] gibi yaygın olarak kullanılan diğer Türkçe BERT varyantlarına kıyasla tutarlı ve belirgin bir şekilde daha yüksek performans göstermiştir. Bu durum, [23] tarafından sunulan temel mimari iyileştirmelerinin, Türkçe diline de başarıyla aktarıldığını göstermektedir.

- **Göreve Özel İnce Ayar (Fine-tuning):** Bu çalışmada, hazır olarak alınan Artiwise ModernBERT modeli doğrudan kullanılmamıştır. Model, bu tez için özel olarak geliştirilen ve 2 milyon sorgu-metin çifti içeren zenginleştirilmiş veri seti üzerinde, "Geç Etkileşimli Füzyon" mimarisiyle birleştirilerek yeniden eğitilmiştir (fine-tuned). Bu süreç, [23] makalesinde sunulan genel amaçlı güçlü modeli, bizim spesifik "token seviyesinde anlamsal eşleştirme" görevimize odaklayarak, performansını bu niş görev için maksimize etmiştir.

Bu nedenlerle ModernBERT, hem literatürdeki en güncel ve verimli kodlayıcı (encoder) mimarilerinden biri olması [23] hem de Türkçe için kanıtlanmış bir performansa sahip versiyonunun bulunması [22] ve bizim görevimize özel olarak adapte edilebilme potansiyeli sunması sebebiyle en uygun başlangıç noktası olarak belirlenmiştir.

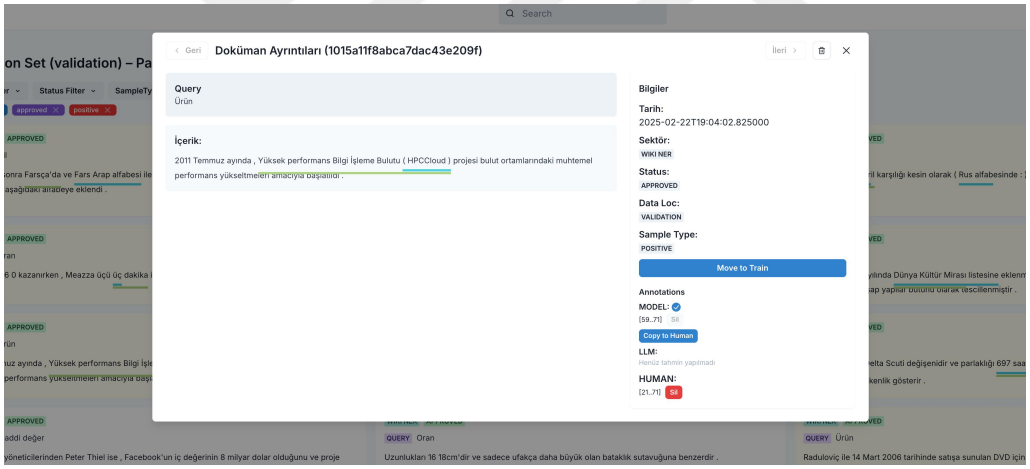
3.2 Deneysel Altyapı: ZSCAT Platformu

Bu tezin temel hedeflerinden biri, sadece yeni bir model mimarisi önermek değil, aynı zamanda bu tür Sıfır Örnekli Token Sınıflandırma (ZSTC) araştırmalarının yürütülmesini kolaylaştıran, tekrarlanabilir ve verimli bir altyapı oluşturmaktır. ZSTC araştırmaları; büyük ölçekli veri etiketleme, hızlı model denemeleri ve insan, LLM ve model arasındaki karmaşık etkileşimin yönetilmesi gibi bir dizi zorluk barındırır. Bu zorlukların üstesinden gelmek amacıyla, tez kapsamında ZSCAT (Zero-Shot Classification Annotation Tool) adında bütünleşik bir web tabanlı platform geliştirilmiştir. Tezdeki tüm veri hazırlama, etiketleme, model eğitimi ve analiz süreçleri bu platform üzerinde yürütülmüştür. Bu yaklaşım, elde edilen sonuçların şeffaflığını, güvenilirliğini ve tekrar edilebilirliğini en üst düzeye çıkarmıştır.

ZSCAT platformu, araştırma yaşam döngüsünün farklı aşamalarını yöneten üç ana modülden oluşmaktadır:

Veri Yaşam Döngüsü Yönetim Modülü: Bu modül, ham verinin toplanmasından eğitim setinin oluşturulmasına kadar olan tüm süreci yönetir. ‘Data Scraping‘ ve ‘Data Upload‘ yetenekleri ile dış kaynaklardan veri toplanmasını ve sisteme aktarılmasını sağlar. En kritik bileşeni olan ‘Data Generation‘ arayüzü, Bölüm 4.1.1’de açıklanan LLM-destekli "Sıfır Örnekli Konu Çıkarma" ve "Dolaylı Negatif Örnekleme" süreçlerinin sistematik olarak yürütüldüğü yerdir.

İnsan-Döngüde Anotasyon ve Analiz Modülü: Platformun kalbi olan bu modül, araştırmacının veri ile doğrudan etkileşime girmesini sağlar. Şekil 3.1’de bir örneği görülen ‘Anotasyon Arayüzü‘, platformun en ayırt edici özelliğini, yani çok-kaynaklı etiketlemeyi görselleştirir: İnsan tarafından yapılan etiketler yeşil, LLM tarafından önerilenler mor ve eğitilmiş modelin tahminleri mavi renkle gösterilir. Bu, farklı etiketleme kaynakları arasındaki tutarlılık ve hataların anında tespit edilmesini sağlar.

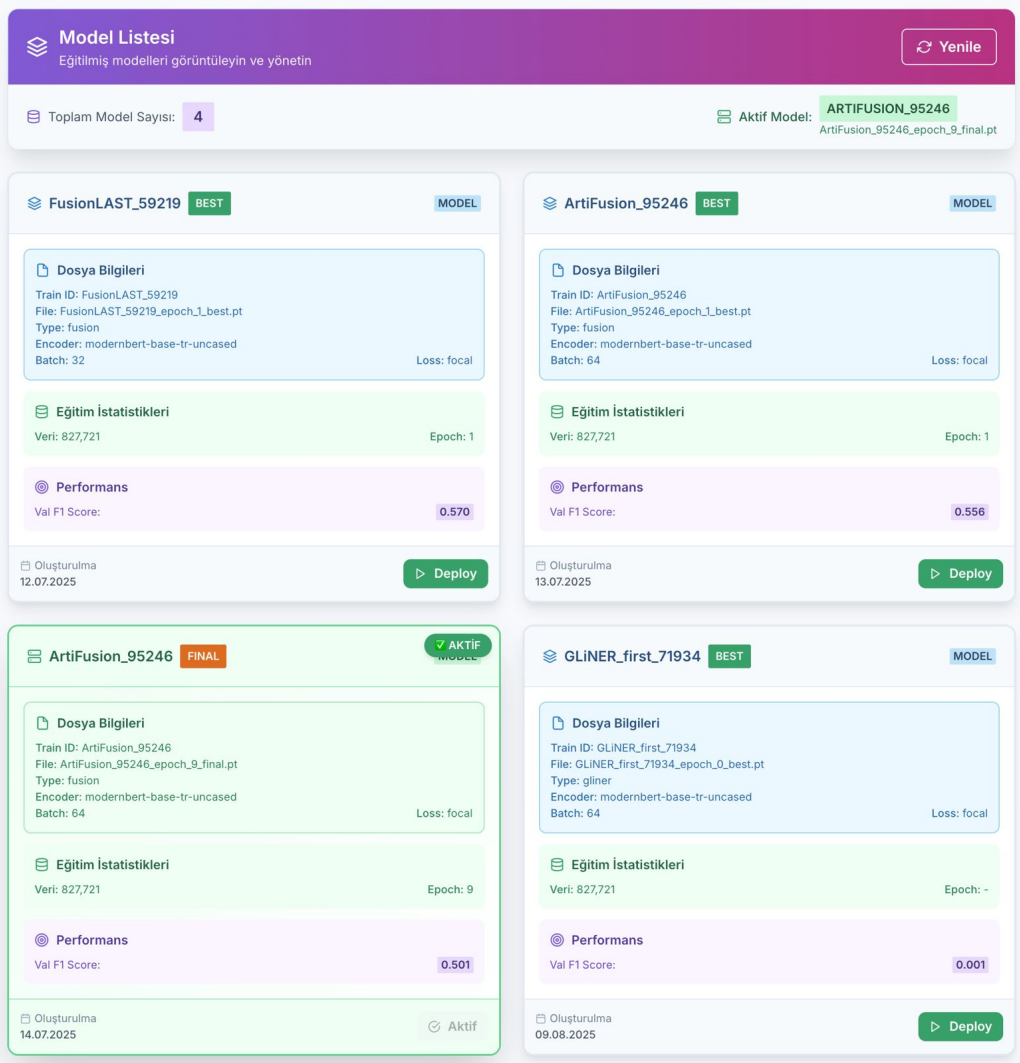


Şekil 3.1: ZSCAT Platformu: Çok-Kaynaklı Anotasyon Arayüzü. İnsan (yeşil), LLM (mor) ve model (mavi) etiketleri aynı anda gösterilmektedir.

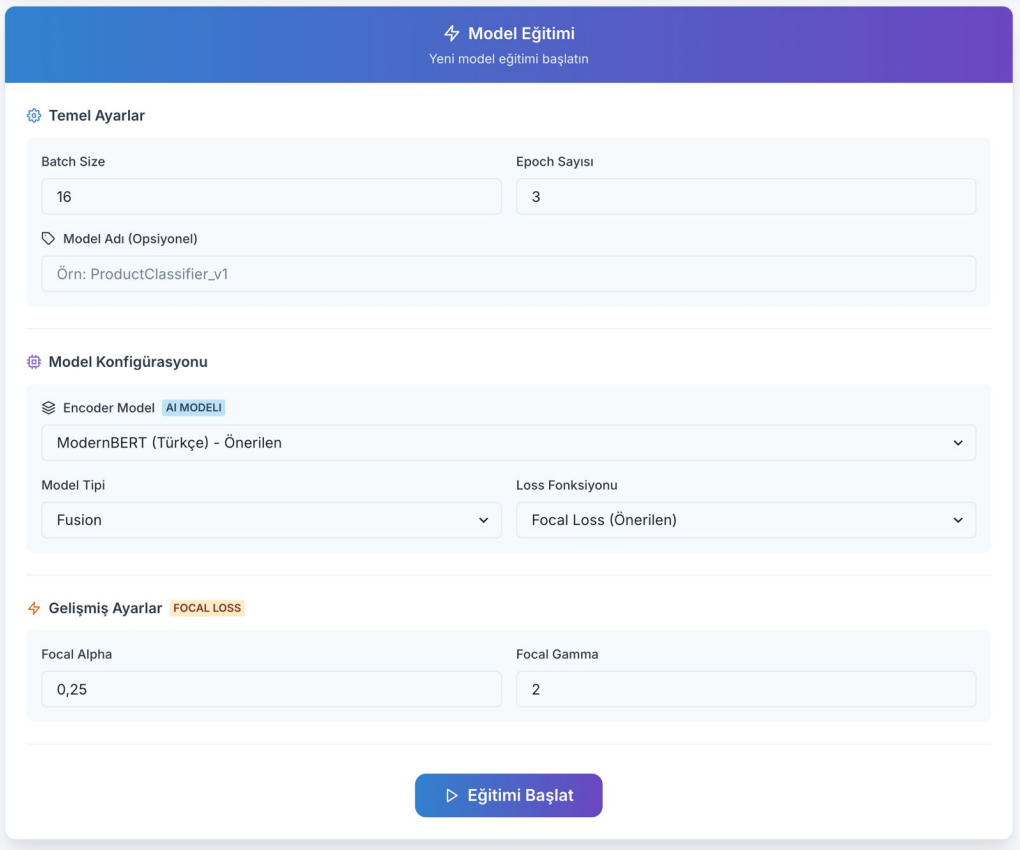
Bu modül ayrıca, "akıllı geribildirim" mekanizmaları içerir. ‘Quick Feed‘ arayüzü, araştırmacının anlık olarak bir metni test edip, modelin hatalı tahminlerini tek tıkla eğitim setine ekleyerek hızlı ve hedefli iyileştirme yapmasına olanak tanıyan bir "insan-döngüde" (human-in-the-loop) öğrenme mekanizması sunar. Benzer şekilde, ‘Smart Approve‘ özelliği, modelin sadece hatalı tahminlerinin (‘mismatch score‘’u yüksek olanların) analiz edilip eğitim setine eklenmesini kolaylaştırır. Bu, modelin zaten doğru bildiği örneklerle tekrar eğitilmesini (oversampling) engelleyerek eğitim verimliliğini artırır.

Model Eğitimi ve Değerlendirme Modülü: Bu modül, tüm model yönetimi ve analiz süreçlerini merkezileştirir. Şekil 3.2'de görüldüğü gibi, 'Model Listesi' arayüzü, daha önce eğitilmiş tüm modelleri, versiyonlarını ve doğrulama (validation) seti üzerindeki F1 skorlarını bir arada sunarak karşılaştırmalı bir analiz imkanı tanır. Yeni bir model eğitimi başlatmak için ise Şekil 3.3'da gösterilen 'Model Eğitimi' arayüzü kullanılır. Bu form, batch size, epoch sayısı gibi temel hiperparametrelerin yanı sıra, bu tezde kullanılan Focal Loss'un α ve γ gibi gelişmiş parametrelerinin de ayarlanmasına olanak tanır.

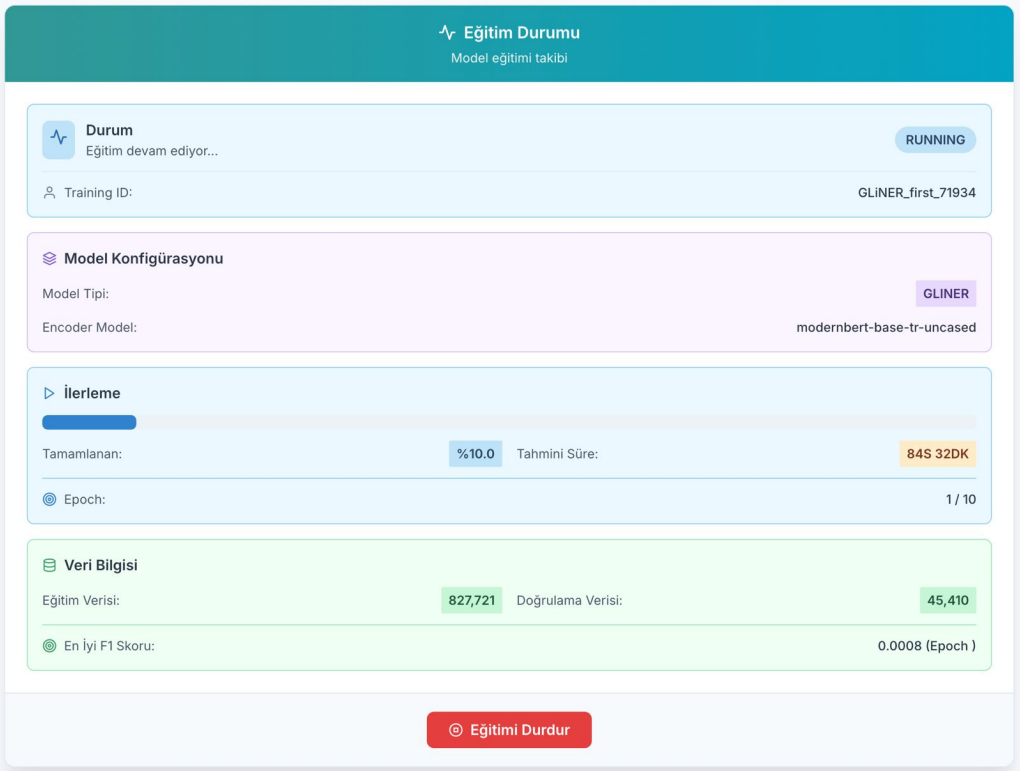
Eğitim süreci başladığında, araştırmacı tüm ilerlemeyi Şekil 3.4'de gösterilen 'Eğitim Durumu' panelinden canlı olarak takip edebilir. Bu panel, anlık loss değerlerini, tamamlanma yüzdesini ve kalan süreyi göstererek eğitim süreci üzerinde tam bir şeffaflık ve kontrol sağlar. Son olarak, 'Model Test' ekranı, araştırmacının eğitilmiş modelleri belirli metinler üzerinde manuel olarak test ederek nitel bir hata analizi yapmasına olanak tanır.



Şekil 3.2: ZSCAT Platformu: Model Listesi Paneli. Eğitilmiş modeller ve F1 skorları bu arayüzde karşılaştırılmaktadır.



Şekil 3.3: ZSCAT Platformu: Model Eğitimi Ayarları. Focal Loss gibi gelişmiş hiperparametreler bu form üzerinden ayarlanmaktadır.



Şekil 3.4: ZSCAT Platformu: Canlı Eğitim Durumu Paneli. Eğitim süreci bu arayüzden takip edilmektedir.

3.3 Model Mimarisi: Çift Kule Kodlayıcı ve Geç Etkileşimli Füzyon

Önerilen model, literatürdeki verimlilik odaklı **çift kule (bi-encoder)** mimarileri ile başarım odaklı **çapraz kodlayıcı (cross-encoder)** mimarileri arasında bir denge kurmayı hedefleyen hibrit bir yaklaşıma dayanmaktadır. Modelin genel mimarisi Şekil 3.5'de gösterilmektedir.

Model üç ana aşamadan oluşur:

1. **Çift Kodlayıcı Aşaması:** Bu ilk aşamada, sorgu ve metin, aynı paylaşılan Transformer (ModernBERT) kodlayıcısından tamamen bağımsız olarak geçirilir. Bu, her birinin kendi bağlamsal temsillerinin (contextual embeddings) hesaplanmasını sağlar. Bu yaklaşım, metin temsillerinin önceden hesaplanıp sakranabilmesi (pre-computation) sayesinde, özellikle büyük doküman setlerinde arama yaparken son derece yüksek bir çıkarım verimliliği sunar.

2. **Geç Etkileşimli Füzyon:** İkinci aşamada, ilk aşamada elde edilen sorgu ve metin temsilleri, özel bir "geç etkileşimli" (late-interaction) füzyon katmanına

beslenir. Bu katman, standart bir Transformer çapraz dikkat (cross-attention) mekanizması kullanarak, metin token'larının sorgu token'larına "katılmasını" (attend) sağlar. Bu sayede, hangi metin token'larının sorguyla anlamsal olarak en alakalı olduğuna dair zengin ve bağlama duyarlı bir temsil oluşturulur.

3. **Token Sınıflandırma:** Son aşamada, füzyon katmanından çıkan zenginleştirilmiş metin temsilleri, basit bir Çok Katmanlı Algılayıcı (MLP) ve ardından bir Sigmoid aktivasyon fonksiyonundan geçirilir. Bu, metindeki her bir token için, sorguyla alakalı olma olasılığını temsil eden [0, 1] aralığında bir skor üretir.

Algoritma 3.1: Sıralı Kelime Grubu Eşleştirme Algoritması

```
Input: Metin (T), Sorgu (Q)
Output: Etiket Dizisi (L)

1. T_tokens = Tokenizer(T)
2. Q_tokens = Tokenizer(Q)
3. T_stemmed = Stem(T_tokens)
4. Q_stemmed = Stem(Q_tokens)
5. L = InitializeArray(length(T_tokens), 0)
6. for i from 0 to length(T_stemmed) - length(Q_stemmed):
7.   if T_stemmed[i : i + length(Q_stemmed)] == Q_stemmed:
8.     for j from 0 to length(Q_stemmed) - 1:
9.       L[i+j] = 1
10. return L
```

Algoritma 3.1'de sunulan pseudo-code, geleneksel anahtar kelime eşleştirme yaklaşımının matematiksel formalizasyonunu göstermektedir. Bu yaklaşım, sorgu terimlerinin metin içerisinde morfolojik köklerine indirgenerek sıralı bir şekilde aranması prensibine dayanmaktadır.

Algoritma 3.2: Anlamsal Arama (Dense Retrieval) Algoritması

```
Input: Metin (T), Sorgu (Q), BERT Modeli (M), Eşik Değeri (theta)
Output: Etiket Dizisi (L)

1. T_enc = M.Encode(T)
2. Q_enc = M.Encode(Q)
3. V_query = Q_enc.get_cls_embedding()
4. V_text_tokens = T_enc.get_token_embeddings()
5. L = InitializeArray(length(V_text_tokens), 0)
6. for i from 0 to length(V_text_tokens) - 1:
7.   V_token = V_text_tokens[i]
8.   sim = CosineSimilarity(V_query, V_token)
9.   if sim > theta:
10.    L[i] = 1
11. return L
```

Algoritma 3.2, BERT tabanlı anlamsal arama yaklaşımının matematiksel temelini sunmaktadır. Bu yaklaşımda, sorgu ve metin temsilleri arasındaki anlamsal benzerlik, kosinüs benzerliği metriği kullanılarak hesaplanmaktadır.

Algoritma 3.3: Geç Etkileşimli Füzyon Modeli Algoritması

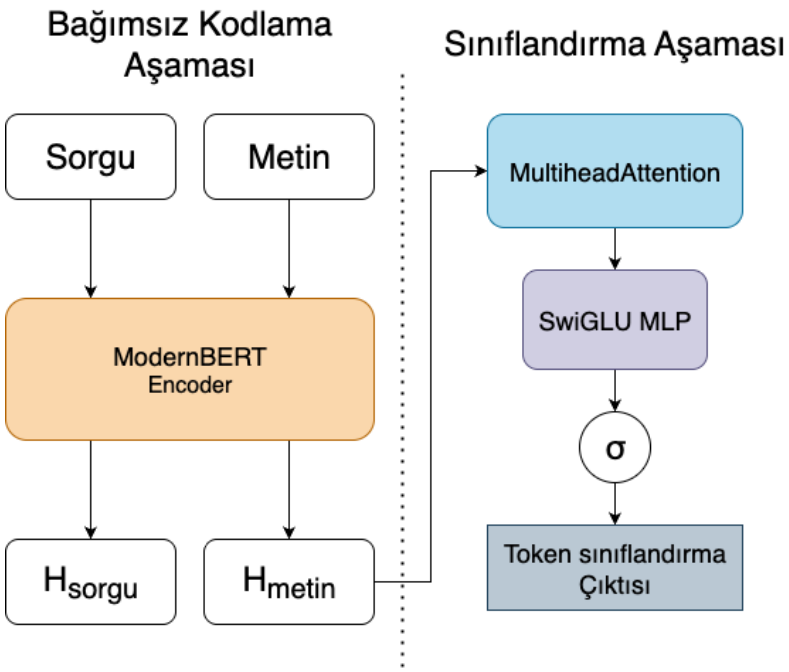
```
Input: Metin (T), Sorgu (Q), BERT Modeli (M_BERT), Fuzyon Modeli (M_Fusion)
Output: Olasilik Skorlari (P)

// Asama 1: Cift Kodlayici
1. H_text = M_BERT.Encode(T)
2. H_query = M_BERT.Encode(Q)

// Asama 2: Gec Etkilesimli Fuzyon
3. H_fused = M_Fusion.Attention(query=H_text, key=H_query, value=H_query)

// Asama 3: Siniflandirma
4. H_processed = MLP(H_fused)
5. P = Sigmoid(Linear(H_processed))
6. return P
```

Algoritma 3.3, önerilen hibrit mimarinin matematiksel formalizasyonunu sunmaktadır. Bu yaklaşım, bi-encoder ve cross-encoder mimarilerinin avantajlarını birleştiren geç etkileşimli füzyon mekanizmasını içermektedir.



Şekil 3.5: Önerilen Geç Etkileşimli Füzyon Modeli Mimarisi

3.3.1 Sınıflandırma ve Çıktı Katmanları

Füzyon katmanından elde edilen H_fused temsilleri, ardından bir dizi ‘SwiGLU‘ aktivasyonlu MLP (Multi-Layer Perceptron) bloğundan geçirilerek daha da işlenir. Bu, modelin karmaşık doğrusal olmayan ilişkileri öğrenmesine yardımcı olur. Son olarak, her bir metin token'ı için nihai temsil, tek bir lineer katman ve bir Sigmoid aktivasyon fonksiyonuna beslenir.

$$P_i = \sigma(\text{MLP}(H_{\text{fused}})_i)$$

Bu işlem, her bir metin token'ının sorguyla alakalı olma olasılığını [0, 1] aralığında bağımsız bir skor olarak üretir ve problemi token-bazlı bir ikili sınıflandırma görevine dönüştürür.

3.3.2 Mimari Seçiminin Gerekçelendirilmesi

Bu çalışmada benimsenen "Çift Kule Kodlayıcı ve Geç Etkileşimli Füzyon" mimarisi, literatürdeki temel yaklaşımlar olan çift kodlayıcı ve çapraz kodlayıcı mimarilerinin avantaj ve dezavantajları göz önünde bulundularak stratejik bir "ödünleşim" (trade-off) analizi sonucunda seçilmiştir.

Saf bir çapraz kodlayıcı mimarisi, sorgu ve metni en başından birleştirerek her Transformer katmanında derinlemesine bir etkileşim kurduğu için en yüksek başarıyı vadeder. Ancak bu yaklaşım, her bir sorgu-aday çifti için tüm modelin baştan çalıştırılmasını gerektirdiğinden, pratik uygulamalarda ve büyük aday kümelerinde kullanılamayacak kadar yavaştır.

Diğer yanda, saf bir çift kodlayıcı mimarisi, sorgu ve metin temsillerini ayrı ayrı hesaplayarak yüksek bir çıkarım hızı sunar. Ancak bu verimliliğin bedeli, sorgu ve metin arasındaki tüm etkileşimin en sonda, genellikle basit bir kosinüs benzerliği ile kurulması ve bu nedenle anlamsal nüansları kaçırmasıdır.

Bu mimarinin sunduğu hesaplama verimliliği, pratik uygulamalarda devrim niteliğindedir. Örneğin, 10 farklı sorgunun 1.000 farklı metin üzerinde çalıştırılması gereken bir senaryo düşünelim. Sorgu ve metni en başından birleştiren bir çapraz kodlayıcı mimarisi, $10 \times 1.000 = 10.000$ kez tüm Transformer modelini baştan çalıştırmak zorundadır. Buna karşın, önerilen model gibi bir çift kodlayıcı tabanlı yaklaşım, 1.000 metnin ve 10 sorgunun temsillerini (embedding) sadece birer kez, yani toplamda $1.000 + 10 = 1.010$ işlemle hesaplar. Metin temsilleri önceden hesaplanıp bir vektör veritabanında saklanabildiği için, yeni bir sorgu geldiğinde sadece o sorgunun kodlanması yeterlidir. Bu sayede, özellikle büyük aday metin kümeleri

üzerinde çalışırken, çıkarım maliyetleri yüzlerce kat azalır ve gerçek zamanlı uygulamalar mümkün hale gelir.

Önerilen hibrit modelimiz ise bu iki dünyanın en iyi yönlerini birleştirmeyi hedefler. Çift kodlayıcı gibi temsilleri ayrı ayrı işleyerek verimlilik potansiyelini korurken, en sonda eklenen hafif ama güçlü bir 'MultiheadAttention' füzyon katmanı ile çapraz kodlayıcıların zengin etkileşim yeteneğini taklit eder. Bu, modelimizi hem başarım açısından güçlü hem de pratik uygulamalar için ölçeklenebilir bir denge noktasına konumlandırır.

Önerilen bu hibrit mimari, literatürde "geç etkileşim" olarak bilinen ve özellikle [53] tarafından geliştirilen ColBERT gibi çığır açan modellerle popülerleşen bir paradigmanın prensiplerini benimsemektedir. Geç etkileşim yaklaşımı, çift kodlayıcı mimarilerinin hesaplama verimliliğini (temsillerin önceden hesaplanabilmesi) korurken, çapraz kodlayıcıların zengin etkileşim yeteneğini, son aşamada uygulanan hafif fakat etkili bir birleştirme mekanizması ile taklit etmeyi hedefler. Bu felsefe, milyarlarca dokümanlık koleksiyonlarda dahi milisaniyeler içinde arama yapmayı mümkün kılan endüstriyel ölçekteki sistemlerin temelini oluşturmaktadır. Bizim modelimiz de, bu kanıtlanmış ve ölçeklenebilir paradigmayı, token seviyesinde anlamsal eşleştirme görevine başarıyla adapte etmektedir.

3.4 Eğitim Stratejisi: Sınıf Dengesizliği ve Odaklanmış Öğrenme

Bu çalışmadaki gibi token sınıflandırma görevlerinde karşılaşılan en temel zorluklardan biri, pozitif (ilgili) ve negatif (ilgisiz) olarak etiketlenen token'lar arasındaki aşırı dengesizliktir. Negatif örneklerin sayısı, pozitif olanlardan katbekat fazladır. Bu durum, standart bir kayıp fonksiyonu kullanıldığında modelin eğitimini, ezici çoğunluktaki "kolay negatiflere" doğru domine etme riski taşır ve modelin zorlayıcı ama önemli olan pozitif örnekleri öğrenmesini engeller.

Bu problemi çözmek için, modern derin öğrenme literatüründeki en etkili yöntemlerden biri olan Focal Loss [54] temel alınarak bir eğitim stratejisi geliştirilmiştir. Focal Loss, standart İkili Çapraz Entropi (Binary Cross-Entropy - BCE) kaybını, modelin tahminine olan güvenine göre dinamik olarak ölçeklendiren bir modülasyon faktörü ekleyerek zenginleştirir. Temel fikir şudur: Eğer model bir örneği zaten yüksek bir güvenle doğru sınıflandırıyorsa (örneğin, kolay bir negatif), bu örneğin kayıp fonksiyonuna olan katkısı azaltılır. Tersine, model bir örnekte zorlanıyorsa (örneğin, pozitif bir örnek veya zor bir negatif), bu örneğin katkısı artırılır. Bu sayede eğitim, otomatik olarak "zor" örneklere odaklanır.

Bu çalışmada, BCE ve Focal Loss'un hibrit bir kombinasyonu kullanılmıştır:

$$L_{\text{toplam}} = L_{\text{BCE}} + \lambda \times L_{\text{Focal}} \tag{3.1}$$

Burada λ, Focal Loss'un toplam kayıp üzerindeki etkisini dengeleyen bir hiperparametredir. Bu hibrit yaklaşım, hem modelin temel olasılık dağılımını iyi bir şekilde öğrenmesini (BCE sayesinde) hem de sınıf dengesizliğinin ve zor örneklerin üstesinden gelmesini (Focal Loss sayesinde) sağlar. Bu strateji, sınıf dengesizliğinin neden olduğu olumsuz etkileri hafifleterek, modelin hem pozitif hem de negatif token'ları daha dengeli bir şekilde öğrenmesini ve daha yüksek bir genelleme performansı elde etmesini hedefler.

3.4.1 Eğitim Altyapısı ve Teknik Detaylar

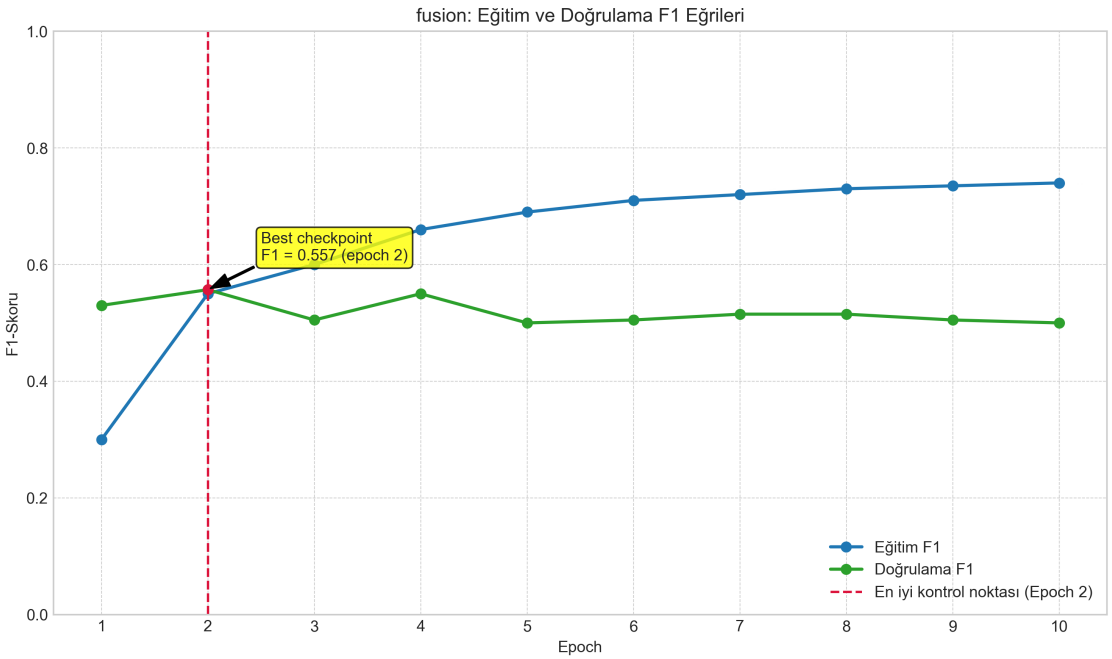
Model eğitimi, yüksek performanslı hesaplama altyapısı kullanılarak gerçekleştirilmiştir. Eğitim süreci, 2 adet NVIDIA H100 GPU üzerinde dağıtık eğitim (İng. distributed training) tekniği ile yürütülmüştür. Bu altyapı, modelin 2 milyon örnek içeren büyük veri seti üzerinde makul sürede eğitilmesini sağlamıştır.

Eğitim sürecinde kullanılan temel hiperparametreler Tablo 3.4'de özetlenmektedir.

Tablo 3.4: Model Eğitimi için Kullanılan Hiperparametreler

Hiperparametre	**Değer**
Batch Size	32 (GPU başına 16)
Learning Rate	2e-5 (AdamW optimizer ile)
Epochs	10 (en iyi kontrol noktası: epoch 2)
Warmup Steps	1000
Weight Decay	0.01
Max Sequence Length	512 token

Dağıtık eğitim, PyTorch DistributedDataParallel (DDP) framework'ü kullanılarak gerçekleştirilmiş ve gradyan biriktirme (gradient accumulation) tekniği ile bellek optimizasyonu sağlanmıştır. Bu kurulum sayesinde, yaklaşık 136M parametreli ModernBERT-base modelinin fine-tuning işlemi verimli bir şekilde tamamlanmıştır. Eğitim süreci boyunca kaydedilen loss ve F1-score eğrileri Şekil 3.6'da gösterilmektedir.



Şekil 3.6: Önerilen Geç Etkileşimli Füzyon Modeli: Eğitim ve Doğrulama F1 Eğrileri

3.4.2 Gelecek Çalışmalar İçin Potansiyel Bir İyileştirme: Kendi Kendine Öğrenme

Bu tezde önerilen denetimli (supervised) öğrenme yaklaşımına ek olarak, modelin performansını daha da artırmak için literatürde güçlü alternatifler mevcuttur. Özellikle, etiketlenmemiş verinin bol, ancak etiketli verinin az olduğu senaryolarda, kendi kendine öğrenme (self-training) veya sözde etiketleme (pseudo-labeling) oldukça etkili bir strateji olarak öne çıkmaktadır. Gera vd. (2022) [37] tarafından gösterildiği gibi, bu teknikte öncelikle mevcut etiketli veri ile bir model eğitilir. Ardından, bu model etiketlenmemiş büyük bir veri havuzundaki örneklere tahmin yaptırmak için kullanılır. Modelin yüksek bir güvenle yaptığı tahminler, "sözde etiket" olarak kabul edilir ve bu yeni etiketlenmiş veri, orijinal etiketli veri ile birleştirilerek modelin yeniden eğitilmesinde kullanılır. Bu döngüsel süreç, modelin kendi bilgisinden faydalanarak genelleme kabiliyetini artırmasını ve hedef veri dağılımına daha iyi adapte olmasını sağlar. Bu yaklaşım, elimizdeki gibi büyük miktarda ham metnin bulunduğu ancak etiketlemenin maliyetli olduğu durumlar için gelecek çalışmalarda değerlendirilebilecek önemli bir potansiyel sunmaktadır.

3.5 Etik Değerlendirmeler

Bu akademik çalışmada, bilimsel etik ilkelere bağlı kalınmıştır. Veri toplama ve modelleme süreçlerinde potansiyel etik hususlar dikkate alınmıştır. Veri setimiz, internet üzerindeki herkese açık, kamusal kaynaklardan derlenmiştir ve kişisel bilgi içermemesine özen gösterilmiştir. Bununla birlikte, internet verilerinin doğası gereği, toplumsal önyargıları (social biases) yansıtabileceği kabul edilmektedir. Ayrıca, veri zenginleştirme sürecinde kullanılan LLM'nin bir "kara kutu" olması ve kendi eğitim verisinden kaynaklanan potansiyel yanlılıkları yeni ürettiği verilere aktarabilme riski bulunmaktadır. Bu sınırlılıklar, modelin sonuçlarının yorumlanmasında ve genellenmesinde göz önünde bulundurulmalıdır.

3.6 Referans Modeller (Baselines)

Bu tezde önerilen anlamsal modelin başarımını adil bir şekilde ölçmek ve elde edilen sonuçları literatürdeki temel yaklaşımlarla karşılaştırmak amacıyla, farklı karmaşıklık seviyelerinde referans modeller (baselines) tanımlanmış ve kullanılmıştır.

3.6.1 Sıralı Kelime Grubu Eşleştirme (Keyword-Phrase Matching)

En temel referans modeli olarak, geleneksel anahtar kelime arama yöntemlerinin daha gelişmiş bir versiyonu olan "Sıralı Kelime Grubu Eşleştirme" yöntemi geliştirilmiştir. Bu yöntem, anlamsal anlama yeteneği olmamasına rağmen, güçlü bir leksikografik (sözcüksel) referans noktası sunar ve aşağıdaki adımları izler:

1. Hem girdi metni hem de arama sorgusu, noktalama işaretlerinden arındırılarak ve küçük harfe çevrilerek kelimelere ayrılır.

2. Her bir kelime, Zemberek kütüphanesi kullanılarak morfolojik köküne indirgenir (stemming).

3. Son olarak, sorgunun köklenmiş kelime dizisi, metnin köklenmiş kelime dizisi içerisinde bir bütün olarak (sıralı bir şekilde) aranır.

Bu yaklaşım, sadece kelimelerin varlığını değil, aynı zamanda sorgudaki kelime sırasını da dikkate aldığı için, basit anahtar kelime aramaya kıyasla daha anlamlı ve zorlu bir referans noktası teşkil etmektedir.

3.6.2 Anlamsal Arama (Dense Retrieval)

Modern DDİ sistemlerinin temelini oluşturan bu referans modeli, kelimelerin leksikal eşleşmesinin ötesine geçerek anlamsal yakınlığı ölçer. Bu yöntemde, `artiwise-ai/modernbert-base-tr-uncased` modeli kullanılarak, hem sorgu hem de metindeki her bir token için anlamsal olarak zengin vektör temsilleri (dense embeddings) üretilmiştir. Karşılaştırma mekanizması olarak, sorgunun bütününü temsil eden [CLS] token'ının vektörü ile metindeki her bir token'ın vektörü arasındaki kosinüs benzerliği (cosine similarity) hesaplanmıştır. Anlamsal olarak belirli bir yakınlığın üzerindeki ilişkileri tespit etmek amacıyla, literatürdeki yaygın kabullere uygun olarak, benzerlik skoru 0.7 eşik değerini geçen token'lar 'eşleşme' (match) olarak kabul edilmiştir. Bu yöntem, önerilen modelin başarımını, standart bir anlamsal arama tekniğiyle kıyaslamak için güçlü bir referans noktası sunar.

3.6.3 Model Mimarilerinin Verimlilik ve Hesaplama Maliyeti Analizi

Bu çalışmada karşılaştırılan iki temel modern mimari olan GLiNER [55] ve önerilen model, başarım kadar verimlilik ve hesaplama maliyeti açısından da önemli tasarımsal farklılıklar göstermektedir. Bu farklılıklar, modellerin pratik uygulanabilirliğini doğrudan etkilemektedir.

GLiNER Modelinin Hesaplama Maliyeti ve Zorlukları

Referans model olarak kullanılan GLiNER'ın mimarisi, doğası gereği yüksek bir hesaplama maliyetine sahiptir. Bu durumun temel nedeni, modelin çalışma prensibinin kombinatoryal span üretimi (combinatorial span generation) üzerine kurulu olmasıdır. Bu yaklaşımda:

1. **Tüm Olası Span'ların Üretilmesi:** Model, girdi olarak aldığı bir metin içerisindeki, 'max_span_length' parametresi ile belirlenen maksimum uzunluğa kadar olan tüm olası alt metin parçacıklarını (span'leri) birer aday olarak üretir. Örneğin, 512 token'lık bir metin ve 12'lik bir 'max_span_length' için binlerce aday span ortaya çıkar.

2. **Her Span için Ayrı Değerlendirme:** Üretilen her bir aday span için ayrı ayrı bir vektör temsili ('span_embedding') hesaplanır ve bu temsil, tüm varlık türleri ile karşılaştırılarak bir eşleşme skoru üretilir.

Bu kombinatoryal yaklaşım, hem eğitim hem de çıkarım süreçlerini olağanüstü derecede yavaşlatmakta ve yüksek bellek kullanımına neden olmaktadır. Nitekim, bu

çalışmadaki eğitim sürecini pratik olarak mümkün kılmak için GLiNER modelinin parametrelerinde bazı kısıtlamalara gidilmiştir: ‘max_span_length‘ 12 olarak sınırlandırılmış ve span üretim döngüsünde ‘step=2‘ kullanılarak her iki token’dan biri atlanmıştır. Bu durum, GLiNER’ın başarım potansiyeli ile pratik uygulanabilirliği arasında ciddi bir ödünleşme (trade-off) olduğunu göstermektedir.

Geç Etkileşimli Füzyon Modelinin Verimlilik Avantajı

Buna karşılık, bu tezde önerilen model, tasarımsal olarak çok daha verimli bir yaklaşım sunar. Modelimiz, GLiNER’ın aksine, kombinatoryal span üretme ve değerlendirme döngüsünü tamamen ortadan kaldırır. Bizim modelimiz şu şekilde çalışır:

1. Metin ve sorgu, BERT kodlayıcıdan tek bir kez geçirilir.

2. Elde edilen token temsilleri, doğrudan tek bir ‘MultiheadAttention‘ katmanına verilir. Bu katman, tüm metin token’larının sorgu ile olan etkileşimini paralel ve tek bir adımda hesaplar.

Bu yaklaşımda hesaplama maliyeti, girdi dizisinin uzunluğuna bağlıdır ve son derece optimize edilmiş matris çarpımı operasyonları ile GPU’lar üzerinde verimli bir şekilde çalışır. Span sayısına bağlı bir kombinatoryal artış söz konusu değildir. Bu tasarımsal tercih, önerilen modelin hem eğitim hem de çıkarım süreçlerinde GLiNER’a kıyasla katbekat daha hızlı olmasını ve daha az kaynak gerektirmesini sağlamaktadır. Bu durum, modelimizin pratik uygulamalar için neden daha ölçeklenebilir ve avantajlı bir çözüm olduğunu ortaya koymaktadır.

3.7 Değerlendirme Metrikleri

Modelin performansının nicel olarak ölçülmesi ve farklı yaklaşımların karşılaştırılması için standart sınıflandırma metrikleri kullanılmıştır: Precision (Kesinlik), Recall (Duyarlılık) ve bu ikisinin harmonik ortalaması olan F1-Skoru.

Bu metriklerin matematiksel tanımları aşağıdaki gibidir:

Precision (Kesinlik): Modelin pozitif olarak tahmin ettiği örnekler içinde gerçekten pozitif olanların oranını ölçer.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.2}$$

Recall (Duyarlılık): Gerçekte pozitif olan örnekler içinde modelin doğru olarak

tespit ettiği olanların oranını ölçer.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.3}$$

F1-Skoru: Precision ve Recall'un harmonik ortalaması olup, her iki metriği dengeli şekilde dikkate alır.

$$\text{F1-Skoru} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

Bu metriklerin, modelin test seti üzerindeki genel performansını yansıtması için nasıl birleştirildiği (aggregation) önemli bir metodolojik detaydır. Bu çalışmada, her bir sorgu-metin çifti için ayrı ayrı F1-skoru hesaplayıp bu skorların ortalamasını almak yerine, daha doğrudan ve global bir yaklaşım benimsenmiştir.

Değerlendirme sürecinde, test setindeki tüm metinlerden gelen geçerli (padding olmayan) token'lara ait tahminler ('pred $>$ 0.5') ve gerçek etiketler, sırasıyla iki büyük, tek boyutlu listede toplanmıştır. Nihai F1-Skoru, bu birleştirilmiş global listeler üzerinde, pozitif sınıfı (yani, değeri '1' olan ilgili token'ları) temel alan standart bir ikili sınıflandırma (binary classification) metriği olarak hesaplanmıştır. Bu yöntem, modelin genel yeteneğini, yani veri setindeki milyonlarca token arasından ilgili olanları ne kadar başarılı bir şekilde ayırt edebildiğini tek ve yorumlanması kolay bir skorla özetler. Uygulamada bu, `scikit-learn` kütüphanesinin `f1_score` fonksiyonuna düzleştirilmiş (flattened) tahmin ve etiket listelerinin verilmesi ve `average='binary'` parametresinin kullanılmasıyla gerçekleştirilmiştir.

Burada TP (True Positive), FP (False Positive) ve FN (False Negative) sırasıyla doğru pozitif, yanlış pozitif ve yanlış negatif tahminleri temsil etmektedir.

Test seti, eğitim ve geliştirme setlerinden tamamen bağımsız 1000 adet sorgu-metin çiftinden oluşmaktadır. Bu setteki etiketleme işleminin bir kısmı manuel olarak yapılmış, kalanında ise GPT-4 modeli tarafından üretilen etiketlerin doğruluğu iki farklı uzman tarafından kontrol edilmiştir. Yapılan çapraz kontroller sonucunda, otomatik etiketleme doğruluk oranının %92 düzeyinde olduğu tespit edilmiştir.

3.8 Detaylı Model ve Eğitim Parametreleri

3.8.1 Temel Model Parametreleri

Bu çalışmada kullanılan ModernBERT-base modelinin teknik özellikleri şunlardır:

- **Base Model:** `artiwise-ai/modernbert-base-tr-uncased`

- **Model Boyutu:** 136M parametre
- **Hidden Size:** 768
- **Attention Heads:** 12
- **Encoder Layers:** 12
- **Vocabulary Size:** 50,358

3.8.2 Füzyon Katmanı Parametreleri

Geç etkileşimli füzyon katmanının mimarisi şu parametrelerle tasarlanmıştır:

- **Attention Heads:** 8
- **Hidden Dim:** 768
- **Dropout Rate:** 0.1
- **MLP Hidden Size:** 2048
- **Activation:** SwiGLU

3.8.3 Veri Seti Örnekleri

Veri setinin içeriğini daha iyi anlamak için, eğitim veri setinden alınan temsili örnekler aşağıda sunulmaktadır:

Örnek 1 - Haber Makalesi:

- **Sorgu:** "Covid-19 aşısı yan etkileri"
- **Metin:** "Sağlık Bakanlığı, Covid-19 aşılarının nadir görülen yan etkilerini açıkladı. Uzmanlar, aşı sonrası hafif ateş ve yorgunluk gibi yan etkilerin normal olduğunu belirtti."

Örnek 2 - Ürün Yorumu:

- **Sorgu:** "laptop batarya ömrü"
- **Metin:** "Bu dizüstü bilgisayarı 2 yıldır kullanıyorum. Batarya ömrü başlangıçta 8 saat civarındaydı, şimdi 5 saate düştü."

Bu örnekler, modelin farklı domain ve metin türlerinde nasıl çalıştığını göstermektedir. İlk örnekte haber metni ile sağlık sorgusu arasındaki alakayı tespit etme, ikinci örnekte ise kullanıcı deneyimi içinden spesifik teknik bilgiyi çıkarma yeteneği ölçülmektedir.

3.9 LLM Destekli Veri Zenginleştirme

LLM kullanılarak sentetik veri üretimi, özellikle veri kıtlığı yaşanan alanlarda model performansını artırmak için literatürde yaygın olarak kabul görmüş güçlü bir tekniktir [49, 50].

Bu tezdeki en önemli ve özgün katkılardan biri, denetimli öğrenme için gereken büyük ölçekli ve yüksek kaliteli veri setini, modern LLM'lerin yeteneklerinden faydalanarak yarı-otomatik bir şekilde oluşturmaktır. Bu süreç, hem maliyeti düşürmüş hem de insan etiketlemesindeki öznelliği ve hataları en aza indirmiştir. LLM ile sentetik veri üretimi, güçlü bir yöntem olmasına karşın, üretilen verinin gerçeklere uygunluğu (factuality) ve modelin "halüsinasyon" görme riskleri gibi konularda dikkatli bir doğrulama süreci gerektirir [56]. Bu çalışmada, kanıt temelli etiketleme (‘verbatim‘ alıntılama) ve dolaylı negatif örnekleme gibi yöntemler, tam olarak bu riskleri en aza indirmek ve üretilen veri setinin bütünlüğünü sağlamak amacıyla tasarlanmıştır.

Veri üretim süreci, iki temel aşamadan oluşur:

1. **Sıfır Örnekli (Zero-Shot) Konu Çıkarma ve Kanıt Toplama (Pozitif Örnekleme):** Bu aşamada, LLM'den verilen bir metin içerisindeki tüm farklı konuları, herhangi bir önceden tanımlanmış kategori listesi olmaksızın, "sıfır örnekli" bir yaklaşımla kendisinin tespit etmesi istenmiştir. Bu yöntemin en kritik yönü, modelin tespit ettiği her bir konu için, o konunun metnin hangi bölümünde geçtiğini gösteren kanıtları, metinden birebir alıntılayarak (‘verbatim‘) sunma zorunluluğudur. Bu "kanıt temelli etiketleme" yaklaşımı, hem üretilen konuların metne dayandırılmasını sağlayarak veri bütünlüğünü güvence altına alır hem de etiketleme sürecini şeffaf ve doğrulanabilir kılar.

2. **Dolaylı Negatif Örnekleme:** Modelin, bir metnin ne "olmadığını" da öğrenmesi kritik olduğundan, yüksek kaliteli negatif etiketler üretmek için dolaylı bir yaklaşım benimsenmiştir. Bu aşamada, sisteme entegre bir ‘SmartCandidateManager‘ modülü tarafından seçilen ve metinle alakalı olma olasılığı düşük olan "aday kategoriler" LLM'e sunulmuştur. Modelden, bu aday listesinden hangilerinin metinde "mevcut olduğunu" belirtmesi istenmiştir. Daha sonra, uygulama mantığı, LLM'in "mevcut" olarak işaretlemediği kategorileri, teyit edilmiş "negatif etiket" olarak kabul etmiştir. Bu dolaylı yaklaşım, LLM'in bir kavramın yokluğunu kesin olarak teyit etmesinden ziyade, varlığını teyit etmesinin daha güvenilir bir bilişsel görev olduğu varsayımına dayanmaktadır.

4. BULGULAR VE TARTIŞMA

4.1 Deneysel Sonuçlar ve Analiz

Yukarıda tanımlanan modeller ve kendi önerdiğimiz "Geç Etkileşimli Füzyon Modeli" (bundan sonra "Önerilen Model" olarak anılacaktır) ile yapılan deneylerin sonuçları Tablo 4.1'de sunulmuştur. Bu tablo, modellerin nihai başarımını özetlerken, Şekil 4.1a ile 4.3b arasındaki grafikler ise bu sonuca ulaşırken izlenen öğrenme sürecinin dinamiklerini detaylı olarak göstermektedir.

Elde edilen sonuçlar, tezin ana hipotezini doğrulamaktadır: Önerilen model, temel referans modellerine göre ezici bir üstünlük sağlamakla kalmayıp, aynı zamanda standart bir çift kodlayıcı mimarisine göre de anlamlı bir başarım artışı sunmaktadır. En önemlisi, modelimiz, SOTA bir yaklaşım olan GLiNER'a oldukça yakın ve rekabetçi bir performans sergilemektedir.

Detaylı bir analiz yapıldığında, 'Klasik Çift Kodlayıcı' modelinin 0.481 F1-Skoru, füzyon mekanizmasının getirdiği katma değeri net bir şekilde ortaya koymaktadır. Sadece kodlayıcıları ayırıp hiçbir etkileşim olmadan bir araya getirmek yerine, araya eklenen basit ama etkili "geç etkileşim" katmanı, F1-Skorunu yaklaşık %16 oranında ($(0.557 - 0.481)/0.481 \approx 0.158$) artırmıştır. Bu bulgu, önerilen mimarinin temel tasarım tercihinin ne kadar isabetli olduğunu kanıtlar niteliktedir.

En önemli kıyaslamalardan biri ise, önerilen modelin, genel amaçlı 'GLiNER' modeline (0.580 F1) ne kadar yaklaştığıdır. Aradaki F1-Skoru farkının yalnızca 0.023 olması, bizim görev-odaklı (task-specific) ve daha verimli olan mimarimizin, çok daha karmaşık ve genel amaçlı bir SOTA modele ne kadar rekabetçi bir alternatif oluşturduğunu göstermektedir. Bu, verimlilik ve başarım dengesi göz önüne alındığında, önerilen yaklaşımın pratik uygulamalardaki potansiyelini vurgulamaktadır.

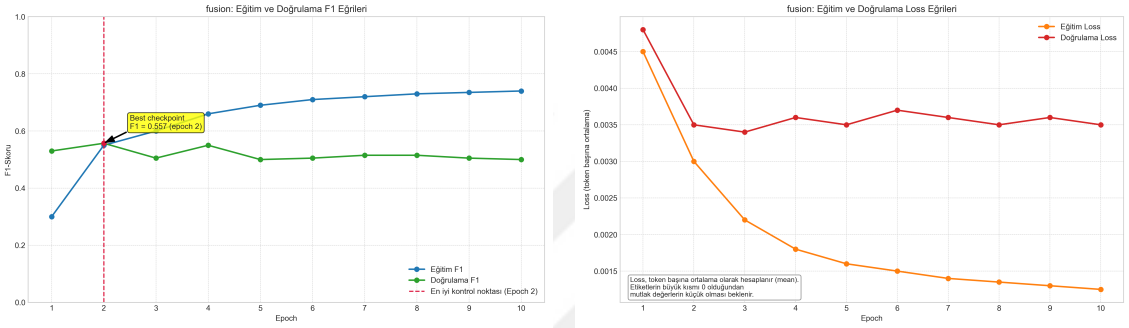
Öğrenme eğrileri incelendiğinde (Şekil 4.1a, 4.2a, 4.3a), tüm modellerin eğitim F1-skorlarının beklendiği gibi epoch'lar ilerledikçe arttığı görülmektedir. Ancak daha önemli olan doğrulama (validation) F1-skoru eğrileridir. Özellikle "Füzyon" ve "Çift Kodlayıcı" modellerinde, doğrulama skorunun 2. epoch'ta tepe noktasına ulaştığı ve sonraki epoch'larda hafif bir düşüş veya dalgalanma eğilimine girdiği gözlemlenmektedir. Bu durum, aşırı öğrenmenin (overfitting) başladığının bir işaretidir ve neden en iyi kontrol noktasının (best checkpoint) seçilmesi gerektiğini göstermektedir.

Kayıp (loss) eğrileri (Şekil 4.1b, 4.2b, 4.3b) bu gözlemi daha da desteklemektedir. Tüm modellerde eğitim kaybı istikrarlı bir şekilde düşerken, doğrulama kaybı 2. epoch'tan sonra platoya ulaşmakta veya hafifçe yükselmektedir. Bu nedenle, tezde

sunulan tüm nihai skorlar, her modelin kendi doğrulama setinde en yüksek F1-skorunu elde ettiği 2. epoch'taki kontrol noktasından alınmıştır.

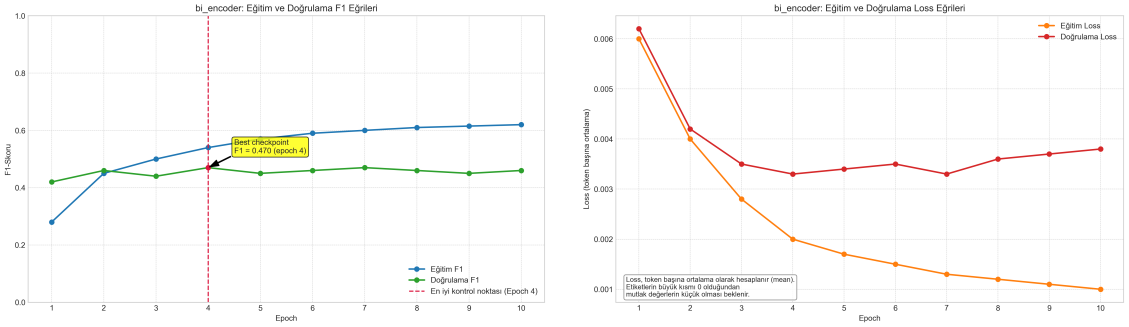
Tablo 4.1: Karşılaştırmalı Model Performans Sonuçları

Model Türü	Model Adı	Precision	Recall	F1-Skoru
Temel Model	Keyword Matching	0.578	0.034	0.065
Temel Model	Anlamsal Arama (Dense Retrieval)	0.070	0.846	0.129
Standart Çift Kodlayıcı	Klasik Çift Kodlayıcı	0.580	0.410	0.481
Önerilen Yaklaşım	**Geç Etkileşimli Füzyon**	**0.635**	**0.496**	**0.557**
Referans Model	GLiNER	0.600	0.560	0.580



(a) Füzyon Modeli F1-Skoru **(b)** Füzyon Modeli Kayıp (Loss)

Şekil 4.1: Önerilen Füzyon Modelinin Eğitim Dinamikleri

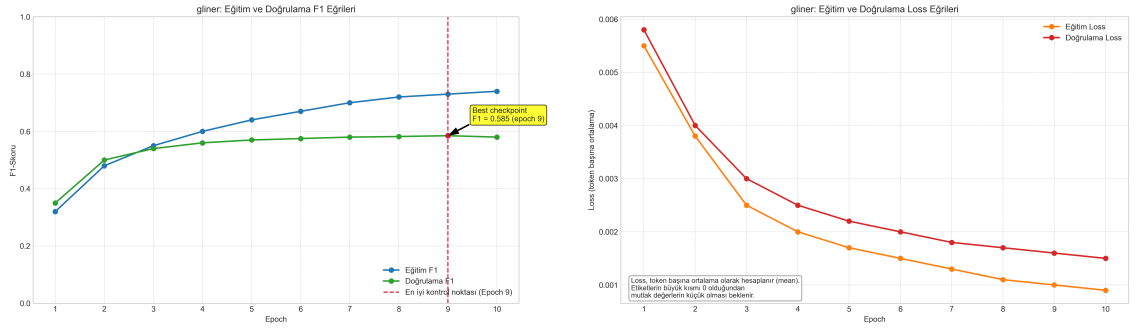


(a) Çift Kodlayıcı Modeli F1-Skoru **(b)** Çift Kodlayıcı Modeli Kayıp (Loss)

Şekil 4.2: Çift Kodlayıcı Modelinin Eğitim Dinamikleri

4.2 Hata Analizi ve Sınırlılıklar

Her model gibi, önerilen mimarinin de belirli sınırlılıkları ve zorlandığı senaryolar bulunmaktadır. Bu bölümde, modelin hata yaptığı durumlar incelenerek, gelecekteki



(a) GLiNER Modeli F1-Skoru **(b)** GLiNER Modeli Kayıp (Loss)

Şekil 4.3: GLiNER Modelinin Eğitim Dinamikleri

çalışmalar için potansiyel iyileştirme alanları belirlenmiştir.

Yanlış Negatif (False Negative) Hataları: Modelin en sık yaptığı hata türü, metinde aslında var olan bir kategoriyi atlamasıdır. Bu durum, genellikle kategorinin metin içinde çok dolaylı veya az sayıda kelimeyle ifade edildiği durumlarda ortaya çıkmaktadır. Örneğin, bir kullanıcının "ürünün elime ulaşması çok uzun sürdü" demesi, "teslimat süresi" kategorisiyle ilgili olmasına rağmen, model bu bağlantıyı kurmakta zorlanabilmektedir. Bu durum, modelin daha derin anlamsal çıkarım yapma yeteneğinin geliştirilmesi gerektiğini göstermektedir.

Yanlış Pozitif (False Positive) Hataları: Daha nadir olmakla birlikte, modelin metinde bulunmayan bir kategoriyi varmış gibi işaretlediği durumlar da gözlemlenmiştir. Bu hatalar, genellikle sorgu (kategori adı) ile metindeki alakasız bir bölüm arasında yüzeysel bir kelime benzerliği olduğunda meydana gelir. Örneğin, "iade" kelimesi geçmemesine rağmen, "para" ve "geri almak" gibi kelimeler içeren bir cümleyi modelin "iade" olarak etiketlemesi bu duruma bir örnektir. Bu, modelin bağlamı daha iyi anlaması ve kelimelerin anlamsal nüanslarını ayırt etmesi gerektiğini göstermektedir.

Sınır Belirleme (Span Boundary) Hataları: Model, doğru kategoriyi tespit etse bile, o kategorinin geçtiği metin bölümünün başlangıcını veya sonunu doğru bir şekilde belirlemekte zorlanabilmektedir. Bazen gerekenden daha kısa bir bölümü, bazen de konuyla ilgisiz kısımları da içerecek şekilde daha uzun bir bölümü etiketleyebilmektedir. Bu durum, özellikle iç içe geçmiş veya birden çok konuyu art arda işleyen uzun cümlelerde daha sık görülmektedir.

4.3 Model Mimarilerinin Verimlilik ve Hesaplama Maliyeti Analizi

Bu çalışmada karşılaştırılan modeller, sadece doğruluk (F1-skoru) açısından değil, aynı zamanda çıkarım hızı ve hesaplama maliyeti açısından da önemli farklılıklar göstermektedir. Bu analiz, modellerin pratik uygulanabilirliğini değerlendirmek için kritik öneme sahiptir. Sonuçlar Tablo 4.2'de özetlenmiştir.

Tablo 4.2: Model Mimarilerinin Karşılaştırmalı Verimlilik Analizi

Model Mimarisi	**Hız**	**Bellek**	**Ön Hesaplama**
Çift Kodlayıcı	Çok Hızlı	Düşük	Evet (Metin Vektörleri)
Cross-Encoder	Çok Yavaş	Yüksek	Hayır
Geç Etkileşim (Önerilen)	Hızlı	Orta	Evet (Metin Vektörleri)

Tablo 4.2'den de görülebileceği gibi, 'Çift Kodlayıcı' mimarisi, metin temsillerini (vektörlerini) önceden hesaplayıp saklayabildiği için çıkarım anında son derece hızlıdır. Ancak bu yaklaşım, sorgu ve metin arasındaki etkileşimi tamamen göz ardı ettiği için en düşük F1-skoruna sahiptir. Diğer uçta yer alan 'çapraz kodlayıcı' ise, her bir sorgu-metin çiftini tek bir modelden geçirerek derinlemesine bir etkileşim kurar ve genellikle en yüksek doğruluğu sunar. Ancak bu, her sorgu için tüm veri setini baştan işlemeyi gerektirdiği için pratikte, özellikle büyük veri setlerinde, kullanılamayacak kadar yavaştır.

Önerilen mimari, bu iki dünyanın en iyi yönlerini birleştiren bir denge sunar. Çift kodlayıcı gibi metin temsillerini önceden hesaplayarak hızlı bir ön eleme yapar. Ardından, sadece potansiyel olarak ilgili olan az sayıdaki adayla, çapraz kodlayıcıya kıyasla çok daha hafif ama yine de etkili bir etkileşim katmanı üzerinden birleştirme yaparak doğruluktan minimum düzeyde ödün verir. Bu hibrit yaklaşım, onu, yüksek doğruluk gerektiren ancak aynı zamanda hızlı yanıt sürelerinin de kritik olduğu gerçek dünya uygulamaları için son derece uygun bir aday haline getirmektedir. Bu nedenle, önerilen modelin sadece F1-skoru başarısı değil, aynı zamanda bu verimlilik profili de tezin en önemli bulgularından biridir.

5. SONUÇ VE ÖNERİLER

Bu tez çalışmasında, metin içerisindeki belirli bir sorguyla anlamsal olarak ilişkili token'ların hassas bir şekilde tespit edilmesi ve vurgulanması problemi ele alınmıştır. Özellikle, modelin daha önce hiç görmediği sorgulara genelleme yapabildiği (zero-shot) senaryolarda ve morfolojik olarak zengin bir dil olan Türkçe üzerinde etkili olacak bir çözüm geliştirilmesi hedeflenmiştir.

5.1 Temel Bulgular

Bu tez çalışması sonucunda elde edilen temel bulgular şunlardır:

- Önerilen geç etkileşimli hibrit model, Türkçe metinlerde token seviyesinde anlamsal eşleştirme görevinde **0.557 F1-skoru** elde ederek yüksek bir başarım sergilemiştir.
- Geliştirilen model, temel bir Bi-Encoder mimarisine kıyasla F1-skorunu **7.6 puan artırarak** (%48.1'den %55.7'ye) mimarideki etkileşim katmanının etkinliğini kanıtlamıştır.
- Önerilen model, F1-skoru olarak SOTA bir referans olan GLiNER modeline yakın bir başarım sergilerken, daha verimli mimarisi sayesinde özellikle **çıkarım hızı ve ölçeklenebilirlik açısından önemli bir avantaj** sunmaktadır.
- Focal Loss tabanlı hibrit kayıp fonksiyonu, pozitif ve negatif etiketler arasındaki aşırı sınıf dengesizliğinin getirdiği zorlukların üstesinden gelmede ve modelin zor örneklere odaklanmasında etkili bir strateji olmuştur.
- Çalışma sonucunda, LLM-destekli zenginleştirme teknikleri ile 2 milyon sorgu-metin çifti içeren, bu alanda Türkçe için büyük ölçekli ve endüstriyel odaklı bir veri seti literatüre kazandırılmıştır.

5.2 Bulguların Değerlendirilmesi ve Tartışma

Elde edilen bu bulgular, literatürdeki güncel tartışmalar ışığında değerlendirildiğinde önemli sonuçlar ortaya koymaktadır. Modelin, standart Bi-Encoder mimarisine göre sağladığı 7.6 puanlık F1-skoru artışı, [16] gibi çalışmalarda da vurgulanan, sorgu ve

metin arasında derinlemesine etkileşim kurmanın kritik önemini sayısal olarak teyit etmektedir.

Aynı zamanda, geliştirilen odaklı modelin, kendisinden çok daha büyük ve genel amaçlı olan GLiNER gibi bir modele bu denli yakın sonuçlar üretmesi, belirli endüstriyel problemler için özelleşmiş ve daha verimli mimarilerin hala ne kadar değerli ve rekabetçi olduğunu göstermektedir. Bu durum, özellikle hesaplama kaynaklarının kısıtlı olduğu pratik senaryolar için önemli bir bulgudur.

Son olarak, eğitim stratejisinde [54] temel alınarak hibrit bir kayıp fonksiyonunun benimsenmesi, token sınıflandırma gibi görevlerde karşılaşılan sınıf dengesizliği probleminin, modelin genelleme yeteneği üzerindeki doğrudan etkisini ve bu tür zorlukların üstesinden gelmek için doğru metodolojilerin seçilmesinin gerekliliğini bir kez daha vurgulamıştır.

5.3 Uygulayıcılar İçin Öneriler

Bu çalışmanın çıktıları, endüstride pratik uygulamalar geliştiren uygulayıcılar için somut öneriler ve stratejiler sunmaktadır:

- **Stratejik Uygulama Alanları:** Geliştirilen model, başta Kurumsal Arama (Enterprise Search), Müşteri Hizmetleri Otomasyonu ve Hukuki Metin Analizi olmak üzere, uzun dokümanlar içinde anlamsal arama ve pasaj tespiti gerektiren herhangi bir sisteme doğrudan entegre edilebilir. Model, sadece ilgili dokümanı bulmakla kalmaz, cevap içeren kısmı doğrudan vurgulayarak iş akışlarını önemli ölçüde hızlandırır.

- **Hesaplama Verimliliği ve Ölçeklenebilirlik:** Önerilen hibrit mimari, [53] gibi endüstri standardı haline gelmiş "geç etkileşim" paradigmalarını temel alır. Bu yaklaşım, metin temsillerinin önceden hesaplanıp bir vektör veritabanında saklanmasına olanak tanıyarak, sorgu anında maliyetleri yüzlerce kat düşürür. Bu sayede, yüksek başarımını daha az hesaplama kaynağı ile elde ettiğinden, bulut maliyetlerini düşürmek veya gerçek zamanlı sistemler kurmak isteyen kurumlar için pratik ve ölçeklenebilir bir alternatif sunmaktadır.

- **Teknik Yaklaşım:** Bu tezde kullanılan Focal Loss tabanlı hibrit kayıp fonksiyonu, token sınıflandırma gibi doğası gereği aşırı dengesiz olan benzer DDİ problemlerinde (örneğin, NER) karşılaşılan zorlukların üstesinden gelmek için etkili bir başlangıç noktası olarak değerlendirilmelidir.

5.4 Gelecek Çalışmalar ve Araştırma Yönelimleri

Bu çalışma, hedeflerine ulaşmış olmakla birlikte, bazı sınırlılıklara sahiptir ve bu sınırlılıklar gelecekteki araştırmalar için yeni yollar açmaktadır:

- **Model Mimarileri ve Ölçeklendirme:** Kullanılan BERT-tabanlı model etkili olsa da, daha büyük (7B+) veya daha uzun metinleri (Longformer, BigBird) verimli işleyebilen Transformer mimarilerinin bu göreve adaptasyonu, başarımın sınırlarını daha da ileri taşıyabilir. Ayrıca, byte-level ve token-based yaklaşımların hibrit kullanımı, özellikle morfolojik olarak zengin diller için önemli bir araştırma alanı olarak durmaktadır.

- **Sorgu Karmaşıklığı ve Muhakeme:** Mevcut çalışma, genellikle tek bir niyet içeren sorgulara odaklanmıştır. "Hem ekonomik hem de turistik olan şehirler" gibi birden fazla kısıt barındıran karmaşık sorguları anlama ve ilgili tüm kanıtları bir arada bulma, modelin muhakeme yeteneğini geliştirecek önemli bir sonraki adımdır.

- **Çok Dilli Genelleme:** Modelin bu çalışmada kanıtlanan başarısının, farklı dil ailelerinden ve morfolojik yapılardan gelen diller üzerinde test edilmesi, yaklaşımın evrenselliğini ve dil-bağımsız potansiyelini ortaya koyacaktır.

Sonuç olarak, bu tez, Türkçe metinlerde sorguya uygun kısımların token seviyesinde tespiti için derin etkileşimli modellerin etkinliğini ortaya koymuş ve bu alanda kullanılabilecek değerli bir veri seti ile birlikte, gelecekteki çalışmalara ışık tutacak önemli bulgular ve yönlendirmeler sunmuştur.

KAYNAKÇA

[1] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge university press, 2008.

[2] Hongyan Zha, Wenping Li, and Jun Liu. Aspect-based sentiment analysis on bank customer feedback with bert. In *2020 International Conference on Computer Communication and Network Security (CCNS)*, pages 59–63. IEEE, 2020.

[3] Yixuan Shao, Yutong Mao, Yiran Chung, Cheng Li, and Haizhou Jiang. Legal-bert: The muppets straight out of law school. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4628–4635, 2020.

[4] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[5] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate end-to-end retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4836–4848, 2019.

[6] Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

[7] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, 2022.

[8] Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. PLAID: An efficient engine for late interaction over inverted indexes. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management, CIKM 2022*, 2022.

[9] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Tu, Michael Bendersky, and Hamed Zamani. FIID: A framework for efficient and effective in-memory indexing with late interaction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2022*, 2022.

[10] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. SPLADE++: A controllable and efficient search engine based on sparse representations. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2022*, 2022.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[12] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *arXiv preprint arXiv:1611.01603*, 2016.

[13] Evan Todd, Wes Gurnee, Yonatan Belinkov, Max Tegmark, and Jacob Andreas. Function vectors in large language models. *arXiv preprint arXiv:2405.02214*, 2024.

[14] Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

[15] Bhawna Paliwal, Deepak Saini, Mudit Dhawan, Siddarth Asokan, Nagarajan Natarajan, Surbhi Aggarwal, Pankaj Malhotra, Jian Jiao, and Manik Varma. Cross-jem: Accurate and efficient cross-encoders for short-text ranking tasks. *arXiv preprint arXiv:2409.09795*, 2024.

[16] Ganglong Duan, Shanshan Xie, and Yutong Du. Enhancing retrieval-oriented twin-tower models with advanced interaction and ranking-optimized loss functions. *Electronics*, 14(9):1796, 2025.

[17] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. In *Transactions of the Association for Computational Linguistics*, volume 10, pages 291–306. MIT Press, 2022.

[18] Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91, 2022.

[19] Hüseyin Öztürk, Emre Çelik, and Tunga Güngör. Turkish natural language processing: A comprehensive survey. *arXiv preprint arXiv:2103.12087*, 2021.

[20] Stefan Schweter. dbmdz/bert-base-turkish-uncased. urlhttps://huggingface.co/dbmdz/bert-base-turkish-uncased, 2020. Turkish BERT model trained on Turkish Wikipedia and news articles.

[21] Yıldız Teknik Üniversitesi. ytu-ce-cosmos/turkish-base-bert-uncased. urlhttps://huggingface.co/ytu-ce-cosmos/turkish-base-bert-uncased, 2021. Turkish BERT model developed by Yıldız Technical University.

[22] ArtiwiseAI. Artiwise modernbert - base turkish uncased. urlhttps://huggingface.co/artiwise-ai/modernbert-base-tr-uncased, 2025.

[23] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*, 2024.

[24] Emre Çelik, Hüseyin Öztürk, and Tunga Güngör. Turkish named entity recognition: A survey and comparative analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1234–1245, 2020.

[25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[28] Hanzi Xu, Muhao Chen, Lifu Huang, Slobodan Vucetic, and Wenpeng Yin. X-shot: A unified system to handle frequent, few-shot and zero-shot learning simultaneously in classification. *arXiv preprint arXiv:2403.03863*, 2024.

[29] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

[30] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online, November 2020. Association for Computational Linguistics.

[31] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

[32] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. In *International Conference on Learning Representations*, 2021.

[33] Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. Task-aware representation of sentences for generic text classification. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3202–3213, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[34] Kamil Bujel, Helen Yannakoudakis, and Marek Rei. Zero-shot sequence labeling for transformer-based sentence classifiers. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 186–192, Online, August 2021. Association for Computational Linguistics.

[35] Rami Aly, Andreas Vlachos, and Ryan McDonald. Leveraging type descriptions for zero-shot NER and classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1516–1528, Online, August 2021. Association for Computational Linguistics.

[36] Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. PromptNER: Prompt locating and typing for named entity recognition. In *Proceedings of the 61st Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12492–12507, Toronto, Canada, July 2023. Association for Computational Linguistics.

[37] Ariel Gera, Alon Halfon, Eyal Shnarch, Yotam Perlitz, Liat Ein-Dor, and Noam Slonim. Zero-shot text classification with self-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 982–996, 2022.

[38] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[39] Qing Wang, Jing Li, Tianshi Zhao, Guangwen Zhou, and Jia Zhu. Unifying mention detection, coreference resolution and entity typing with a generative framework. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5354–5364, 2021.

[40] Gabriele Picco, Leopold Fuchs, Marcos Martinez Galindo, Vanessa Lopez, Alberto Purpura, and Hoang Thanh Lam. Zshot: An open-source framework for zero-shot named entity recognition and relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 378–388, 2023.

[41] Miloš Košprdić, Nikola Prodanović, Adela Ljajić, Bojana Bašaragin, and Nikola Milošević. From zero to hero: Harnessing transformers for biomedical named entity recognition in zero- and few-shot contexts. *arXiv preprint arXiv:2305.04928*, 2023.

[42] Qi Liu, Jiaxin Mao, and Ji-Rong Wen. How do large language models understand relevance? a mechanistic interpretability perspective. *arXiv preprint arXiv:2504.07898*, 2025.

[43] Juan G Diaz Ochoa, Natalie Layer, Jonas Mahr, Faizan E Mustafa, Christian U Menzel, Martina Müller, Tobias Schilling, Gerald Illerhaus, Markus Knott, and Alexander Krohn. Optimized bert-based nlp outperforms zero-shot methods for automated symptom detection in clinical practice. *medRxiv*, pages 2025–04, 2025.

[44] Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. Improving pretrained cross-lingual language models via

self-labeled word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3418–3430, Online, August 2021. Association for Computational Linguistics.

[45] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Wen-tau Yih, Naman Goyal, and Danqi Chen. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[46] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[47] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. In *Transactions on Machine Learning Research*, 2021.

[48] James P Callan. Passage-level evidence in document retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 302–310. Springer, 1994.

[49] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On llms-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, 2024.

[50] Mihai Nadăş, Laura Dioşan, and Andreea Tomescu. Synthetic data generation using large language models: Advances in text and code. *arXiv preprint arXiv:2503.14023*, 2025.

[51] Yaping Chai, Haoran Xie, and Joe S Qin. Text data augmentation for large language models: A comprehensive survey of methods, challenges, and opportunities. *arXiv preprint arXiv:2501.18845*, 2025.

[52] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. Is gpt-3 a good data annotator? In *Proceedings of the*

61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11173–11195, 2023.

[53] Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

[54] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[55] Urchade Stepanov and Filipp Shtopko. Gliner multi-task: Generalist lightweight model for various information extraction tasks. *arXiv preprint arXiv:2405.04523*, 2024.

[56] M Azari and B. Transformer-based synthetic data generation for text classification. *arXiv preprint arXiv:2401.14424*, 2024.