TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DÖNÜŞTÜRÜCÜ DİL MODELLERİNE ETKİLİ HASSAS AYAR YAPMAK İÇİN VERİ MÜHENDİSLİĞİ YÖNTEMLERİ

YÜKSEK LİSANS TEZİ

Muhammed Said ZENGİN

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Öğretim Üyesi Mücahid KUTLU

MART 2022

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, alıntı yapılan kaynaklara eksiksiz atıf yapıldığını, referansların tam olarak belirtildiğini ve ayrıca bu tezin TOBB ETÜ Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırlandığını bildiririm.

Muhammed Said ZENGİN

ÖZET

Yüksek Lisans Tezi

DÖNÜŞTÜRÜCÜ DİL MODELLERİNE ETKİLİ HASSAS AYAR YAPMAK İÇİN VERİ MÜHENDİSLİĞİ YÖNTEMLERİ

Muhammed Said ZENGİN

TOBB Ekonomi ve Teknoloji Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Öğretim Üyesi Mücahid KUTLU

Tarih: MART 2022

Geleneksel yöntemlerle metinden öznitelik çıkarmak ve bir doğal dil işleme görevini yerine getirmek mümkündür, fakat kısıtlı miktardaki etiketli veriyle cümle yapısı ve kelime vektörleri yeterince öğrenilmediği için model performansı kısıtlı kalmaktadır. Bu sebeple son yıllarda araştırmacılar önceden eğitilmiş dil modellerine hassas ayar yapmayı, geleneksel yöntemlere göre daha çok tercih etmektedir. Büyük miktarda veriyle hazırlanan önceden eğitilmiş dönüştürücü dil modelini kullanarak belirli bir görev üzerinde hassas ayar yapmak birçok doğal dil işleme görevinde en yüksek performansı vermektedir. Etiketli veri hazırlamak maliyetli bir işlem olduğu ve etiketli veri sınırlı bir kaynak olduğu için araştırmacılar az veri kullanarak daha iyi sonuç alma yöntemlerini incelemektedir. Bu sebeple veri arttırma yöntemleri olarak aktif öğrenme ve zayıf denetim yolları kullanılmıştır. Aynı zamanda yarı denetimli ve denetimsiz yöntemler de üzerinde çalışılan araştırma konuları olmuştur. Bu tezin kapsamı ise, kısıtlı miktardaki etiketli veri kullanılarak, dönüştürücü dil modellerine en etkili hassas ayar yapma yöntemini araştırmaktır. Etkili hassas ayar yapmak için çapraz dilli eğitim, zayıf denetim, geri çeviri, aşırı örnekleme, aktif öğrenme gibi veri mühendisliği yöntemleri kullanılmıştır. Bu tez kapsamında incelenen konu üç farklı doğal dil işleme görevi üzerinde incelenmiştir. Bu görevler, kontrole değer iddiaların tespiti, taraf tespiti ve konum tespitidir.

Anahtar Kelimeler: Doğal dil işleme, Dönüştürücü dil modelleri, Veri mühendisliği

ABSTRACT

Master of Science

DATA ENGINEERING METHODS FOR EFFECTIVE FINE TUNING TRANSFORMERS LANGUAGE MODELS

Muhammed Said ZENGİN

TOBB University of Economics and Technology
Institute of Natural and Applied Sciences
Department of Computer Engineering

Supervisor: Dr. Öğretim Üyesi Mücahid KUTLU

Date: MARCH 2022

It is possible to extract features from the text and perform a natural language processing task with traditional methods, but the model performance is limited because the sentence structure and word vectors are not learned enough with the limited amount of labeled data. For this reason, in recent years, researchers prefer to fine-tune pre-trained language models more than traditional methods. Fine-tuning a particular task using a pre-trained transformers language model prepared with large amounts of data yields state of the art results in many natural language processing tasks. Because preparing labeled data is a costly process and labeled data is a limited resource, researchers are examining ways to get better results using less data. For this reason, active learning and weak supervision methods were used as data augmentation methods. At the same time, semi-supervised and unsupervised methods have also been studied research topics. The scope of this thesis is to investigate the most effective fine-tuning method for transformers language models using a limited amount of labeled data. Data engineering methods such as cross-language training, weak supervision, back translation, oversampling, active learning have been used for effective fine-tuning. The subject examined in this thesis is detailed on three different natural language processing tasks. These tasks are detecting check-worthy claims, stance detection, and geolocation detection.

Keywords: Natural language processing, Transformer language models, Data engineering

TEŞEKKÜR

Yüksek lisans eğitimim ve tez çalışmalarım boyunca desteğini ve yardımını esirgemeyen, bana sevdiğim bir alanda araştırma imkanı sağlayan değerli hocam Dr. Öğretim Üyesi Mücahid KUTLU 'ya sonsuz teşekkürlerimi sunarım. Bu tezi değerlendiren ve kıymetli görüşlerini paylaşan Doç. Dr. Ahmet Murat ÖZBAYOĞLU ve Prof. Dr. İlyas ÇİÇEKLİ hocalarıma ayrı ayrı teşekkür ederim.

Bu süreçte kıymetli tecrübelerinden faydalandığım TOBB Ekonomi ve Teknoloji Üniversitesi bölümünün değerli öğretim üyelerine, sunduğu burs imkanı ile beni destekleyen TÜBİTAK'a minnettarım.

Birlikte çalışmaktan mutluluk duyduğum Yapay Zeka ve Optimizasyon Birimi çalışma arkadaşlarıma, özellikle de bu zorlu yüksek lisans sürecini başarmayı kolaylaştıran şirketim STM'ye teşekkür ederim.

Son ve en önemli olarak da, hayatımın her döneminde yanımda olan, desteğini hiç esirgemeyen anneme, en büyük yatırımını çocuklarının eğitimine yapan ve her zaman bize yol gösteren babama, değerli fikirleriyle her zaman yanımda olan ablama, zekası ve saygısıyla gurur duyduğum kardeşime gönülden teşekkürlerimi sunarım.

İÇİNDEKİLER

ŞEKİL LİSTESİ

ÇİZELGE LİSTESİ

KISALTMALAR

BERT : Bidirectional Encoder Representations from Transformers
biLM : Deep Bidirectional Language Model
ELMo : Embeddings from Language Models
GCN : Graph Convolutional Networks
GPT : Generative Pretrained Transformer
LSTM : Long Short Term Memory
MLP : Multilayer Perceptron
RNN : Recurrent Neural Networks
TM : Teyide Muhtaç

1. GİRİŞ

İnternet üzerinde hızla artan veri miktarıyla birlikte araştırmacıların veriye erişimi de artmıştır. Teknolojinin ve işlemci gücünün gelişmesiyle birlikte milyonlarca veri kullanarak dönüştürücü dil modeli eğitmek kolaylaşmıştır. Son yıllarda önceden eğitilmiş dönüştürücü dil modellerin eğitilmesi ve açık kaynaklı bir şekilde paylaşımı yaygınlaşmıştır. Erişim kolaylığı sayesinde araştırmacılar doğal dil işleme çalışmalarında oldukça yoğun bir şekilde önceden eğitilmiş dönüştürücü dil modellerine hassas ayar yapmaktadır. Kısıtlı miktardaki etiketli veri üzerinde geleneksel yöntemlerle öznitelik çıkarmak, milyonlarca veri ile eğitilmiş dönüştürücü dil modeli kullanmaya göre eski bir yöntem olarak görülmektedir.

Dönüştürücü dil modeli, dikkat mekanizmasını kullanan bir derin öğrenme modelidir. Tekrarlayan sinir ağlarında olduğu gibi verilen girdileri sıralı bir şekilde işlemek için tasarlanmıştır. Fakat verilerin sırayla işlenmesi gerekmez. Dikkat mekanizması girdi dizisindeki herhangi bir konum için bağlam sağlar. Girdi olarak bir metin verildiğinde, cümlenin başlangıcını sonundan önce işlemesi gerekmez. Bunun yerine, cümledeki her bir kelimeye anlam veren bağlamı tanımlar. Bu özellik, tekrarlayan sinir ağına göre daha fazla paralelleştirmeye izin vermektedir ve eğitim sürelerini azaltmaktadır.

Önceden eğitilen dil modeli, sınıflandırma, varlık ismi tanıma, soru cevaplama gibi doğal dil işleme görevleri için hassas ayar yapılarak kullanılmaktadır. Hassas ayar, önceden eğitilmiş bir sinir ağının ağırlıklarını başlangıç olarak kabul ederek, yeni veriyle bu ağırlıklarının değiştirilmesidir. Normal eğitimden farklı olarak, önceden kullanılan büyük verinin tüm bilgisine sahip olduğu için, hassas ayar yapmak daha düşük veriyle daha başarılı sonuçlar vermektedir.

Hassas ayar yapmak için önceden eğitilmiş bir dönüştürücü dil modeli temel olarak alınmaktadır. Ardından görev için gerekli etiketli veriyle, model parametrelerini ayarlayarak yeniden bir eğitim yapılmaktadır. Ardından modelin son katmanında çıktı olarak istenen görev ayarlanmaktadır. En yüksek performans gösteren çalışmalar, kısıtlı miktarda etiketli veri kullanarak en iyi sonucu almak için hassas ayar yöntemlerini kullanmaktadır. Bu sebeple, kısıtlı eğitim verisi kullanarak en efektif hassas ayar yöntemi önem kazanmıştır.

Bu çalışmada dönüştürücü dil modellerine etkili hassas ayar yapmak için veri mühendisliği yöntemleri üç farklı görev üstünde incelenmiştir. Bu görevlerde incelenen veri mühendisliği yöntemleri, dile özel eğitim, veri dağılımı eşitleme, makine çevirisi, zayıf denetim ve çapraz dilli eğitim yöntemleri olmuştur. İncelenen doğal dil işleme görevleri ise, kontrole değer iddiaların tespiti görevi, taraf tespiti görevi ve konum tespiti görevidir.

Kontrole değer iddiaların tespiti, girdi olarak verilen bir metnin kontrol edilmeye değer bir iddia olup olmadığının incelenmesidir. Sosyal medya platformları, insanlarla kolayca iletişim kurmak için uygun bir ortam sağlamaktadır. Bu nedenle birçok kişi bu platformlarda dilediği mesajı paylaşarak ifade özgürlüğünün tadını çıkarır. Bununla birlikte aynı platformlar toplum üzerinde büyük olumsuz etkisi olan yanlış bilgileri yaymak için de kullanılmaktadır.

Birçok gazeteci, sosyal medyada gördüğü iddiaların doğruluğunu araştırarak doğruluk kontrol sitelerinde bulgularını paylaşmaktadır ve yanlış bilginin yayılmasına karşı mücadele etmektedir. Doğruluk kontrol siteleri yanlış bilgiyle mücadelede hayati öneme sahiptir ve aktif olarak kullanılmaktadır. Fakat yanlış haberler doğru haberlerden daha hızlı yayıldığı için ve doğruluk kontrolü zaman alıcı bir süreç olduğu için sorun sürekli olarak devam etmektedir. Bu nedenle, yanlış bilgilendirmeyle mücadelede doğruluk kontrolü yapan gazetecilere yardımcı olacak sistemler gerekmektedir.

Doğruluk kontrol çalışmalarının nihai hedefi, iddiaların doğruluğunu otomatik olarak tespit eden sistemler oluşturmaktır. Ancak kullanıcılar internette gördükleri her bilgiyi paylaşmaya devam ettiği sürece, "mükemmel" bir doğruluk kontrol sistemi kurmak yanlış bilginin yayılmasını ve bunun olumsuz sonuçlarını engelleyemez. Bu nedenle, sosyal medya kullanıcılarını paylaşım yaptığında uyaran sistemlere de ihtiyaç bulunmaktadır. Kullanıcının paylaştığı metnin kontrol edilmeye değer bir iddia olup olmadığını tespit eden bir sistem gerekmektedir. Bu sistemler, doğruluk kontrolü yapan kişilerin teyit edilecek iddiaları tespit etmelerine yardımcı olarak, çabalarını iddia kontrolü için harcamalarına olanak tanır.

Bu sebeplerden dolayı, araştırmacılar son yıllarda kontrole değer iddiaların tespiti çalışmalarına yönelmiştir.

Taraf tespiti verilen bir metnin belirli bir hedefe karşıt veya destekleyen tarafta olduğunun tespit edilmesini sağlayan bir metin sınıflandırma problemidir. İnsanlık, tarih boyunca birbirine zıt fikirler ortaya atarak çeşitli taraflar oluşturmuştur. İnsanlar kendilerine bir taraf seçmekte ve seçtiği tarafı toplum içinde desteklemektedir. Toplum, bilindiği gibi sanal bir ortama büyük bir hızla taşınmaktadır. Aynı şekilde taraflar, düşünceler, fikirler ve insanlar da sosyal medyada var olmaya çalışmaktadır. Bu sebeple sosyal medyada birçok konu hakkında taraflaşma ve fikir bölünmeleri bulunmaktadır. Konuyu destekleyenler ve karşı olanlar fikirlerini beyan etmektedirler.

Sosyal medya üzerinde taraf tespiti ise birçok alan için büyük öneme sahiptir. Örneğin, anket firmaları sosyal medyanın nabzını ölçerek çeşitli analizler yapmaktadır. Büyük şirketler, reklam verenler, tanınmış kişiler ve siyasi partiler bu tarzdaki verileri inceleyerek gelecek planlarını yapmaktadırlar. Aynı zamanda terör propagandası yapan sosyal medya hesapları ve toplumsal düzeni bozmaya yönelik oluşturulan büyük bot gruplarının tespiti için de taraf tespitinden yararlanılmaktadır. Bu sebeple sosyal medya verisi üzerinde taraf tespiti konusu, üzerinde durulması gereken ve başarılı sonuçlar alınması gereken önemli bir konudur.

Yapılan ilk taraf tespiti çalışmaları konuya özel olmuştur. Fakat her konu için ayrı bir etiketli veri hazırlama problemi ortaya çıkmıştır. Bunu çözmek için çok hedefli ve çapraz hedefli çözümler sunulmuştur. Fakat yine de tüm konular için taraf tespiti yapmak mümkün olmamıştır. Bu sebeple sıfır atışlı çözümler de incelenmiştir.

Konum tespiti bir sosyal medya paylaşımının konumunu tespit etmeyi veya paylaşım yapan kullanıcının yaşadığı yeri tespit etmeyi amaçlayan bir problemdir. Twitter, Facebook ve Tumblr gibi sosyal medya servislerinde üretilen veriler doğal afet müdahalesinden hedefli reklamcılığa kadar çeşitli sektörlerde kullanılmaktadır. Çoğu zaman bu görevler için kullanıcı konumuna ihtiyaç duyulmaktadır. Örneğin reklamcılar, kullanıcı konumuna göre reklamları değiştirebilir, arttırabilir veya azaltabilir. Sosyal medya platformları kullanıcılardan konum bilgisi istese de bu bilgiler geçici ve yapısal olmayan türdedir. Bu sebeple metin tabanlı veya kullanıcının meta bilgisi tabanlı coğrafi konum tespiti araştırmaları yaygınlaşmıştır.

Bu problem için kullanıcının paylaştığı metin verileri, kullanıcı bilgileri ve arkadaşlık bilgileri kullanılmaktadır. Farklı veri kombinasyonlarını kullanan hibrit sistemler de sunulmuştur. Kullanıcı paylaşımları gibi metin verilerinin yanı sıra, kullanıcının saat dilimi, takipçi sayısı ve takipçileri gibi meta veriler de konum tespitinde kullanılmıştır. Takipçi ağı ve bahsetme ağının kullanıldığı hibrit çalışmalar en gelişmiş sonuçları vermektedir. Fakat yüzlerce arkadaş ve bahsetmenin bulunduğu ve tüm meta verilere erişemediğimiz sosyal medya platformlarında konum tespiti kısıtlı kalmaktadır. Metin verisi diğerlerine göre daha kolay elde edildiği için araştırmacılar daha çok metin verisiyle coğrafi konum tespiti yapmaktadır.

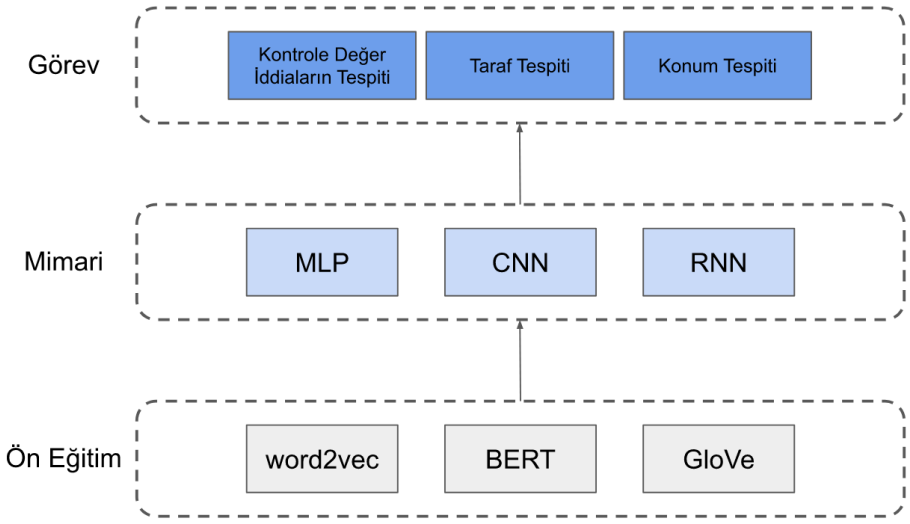
2. ARKAPLAN

Bu bölümde çalışmada geçen konular ve kullanılan algoritmalar alt başlıklar halinde detaylıca açıklanacaktır. Sırasıyla doğal dil işlemede ön eğitim, dönüştürücü dil modelleri ve BERT konuları anlatılacaktır. Ardından kullanılan GCN ve MLP algoritmaları detaylandırılacaktır. Son olarak kullanılan veri detayları, veri çekme yöntemleri ve Twitter hakkında bilgiler bulunmaktadır.

2.1 Doğal Dil İşleme'de Ön Eğitim

Doğal dil işlemede ön eğitim, daha sonra kullanılacak bir görev için modelin önceden eğitilmesidir. Farklı görevler için büyük veri üzerinde tekrar tekrar model eğitmenin maliyetli olmasından dolayı, modeller önceden eğitilip kullanılmaktadır. Bu sayede önceden eğitilmiş model parametreleri kullanılarak, belirli bir görev için daha kısa sürede daha başarılı sonuçlar alınmaktadır. Bu kavram insandan esinlenilmiştir. Önceden eğitilen ve hazırlanan bir insan ileride yapacağı bir görevi daha başarılı bir şekilde yapabilir.

Şekil 2.1'de görüldüğü üzere, önceden eğitilmiş metin temsilleri, farklı doğal dil işleme görevleri için çeşitli derin öğrenme mimarilerinden faydalanmaktadır. Metni anlamak için öncelikle metin temsillerini öğrenmek gerekmektedir. İnternette var olan büyük miktardaki metin verisi kullanılarak kendi kendini denetleyen öğrenme, metin temsillerini önceden eğitmek için yaygın olarak kullanılmıştır. Bu sayede modeller hiçbir etiketlemeye gerek kalmadan devasa metin verilerinden denetim yoluyla öğrenmektedir.



Şekil 2.1: Ön eğitim mimarisi.

Kelime temsilleri, kelimelerin vektör uzayında temsil edilmesi için kullanılan terimdir. Birbiriyle alakalı kelimelerin vektör uzayında yakın olmasını sağlamaktadır. Kelime temsilleri doğal dil işleme için derin öğrenmenin temelini oluşturur. Kelime temsilleri alınan bir metin verisi üzerinden önceden eğitilir ve kaydedilir. Birlikte ortaya çıkma istatistiğini kullanarak oluşturulur. Örneğin "işe araba ile gittim" ve "işe otobüs ile gittim" cümlelerinde geçen araba ve otobüs kelimelerinin vektör uzayında yakın olması beklenmektedir.

Kelime temsilleri araştırmalarda aktif olarak kullanılmaktadır. Fakat çok büyük veriler üzerinde vektörler oluşturulduğu zaman bağlamsal temsiller kaybedilmektedir. Genel bir veri üzerinde çıkarılan kelime vektörleri özel bir domainde kullanılamaz. Çünkü eş sesli kelimeler temsil edilemez ve anlam kaybı olmaktadır. Bu sebeple bağlamsal temsiller kullanılır. Bağlamsal temsillerde eş sesli iki kelime, iki farklı vektöre sahiptir.

Bu sebeple kelime temsil yöntemi kullanmak yerine 2015 yılında Dai ve diğ. [15] LSTM tabanlı bir dil modeli üretilmiş ve bağlamsal temsilleri de dikkate almıştır. LSTM sayesinde model kelime dizilerini ve cümleleri bağlamsal olarak öğrenebilmiştir. Yarı denetimli sinir ağları ile öncelikle bir dil modeli oluşturulmaktadır, ardından spesifik bir doğal dil işleme görevi için hassas ayar yapılmaktadır. Ardından 2017 yılında yayınlanan ELMo [60] kelime vektörlerini, büyük bir metin bütünü üzerinde önceden eğitilmiş bir derin çift yönlü dil modelinin (biLM) iç durumlarının öğrenilmiş işlemi olarak temsil etmiştir. Temel olarak sözdizimi ve anlambilim gibi kelimelerin karmaşık özelliklerini ve bu kullanımların dilsel bağlamlar arasında nasıl değiştiğini incelemişlerdir. Üçüncü olarak OpenAI araştırma şirketi tarafından 2018 yılında yayınlanan GPT [62] çalışması doğal dil işleme görevleri için dönüştürücü tabanlı bir mimari ve eğitim prosedürü sunmuştur. Eğitim iki aşamadan oluşur. İlk olarak, bir sinir ağı modelinin başlangıç parametrelerini öğrenmek için etiketlenmemiş veriler üzerinde bir dil modelleme hedefi kullanılır. Daha sonra, bu parametreler sınıflandırma gibi bir hedef göreve uyarlanır.

2.2 Dönüştürücü Dil Modelleri ve BERT

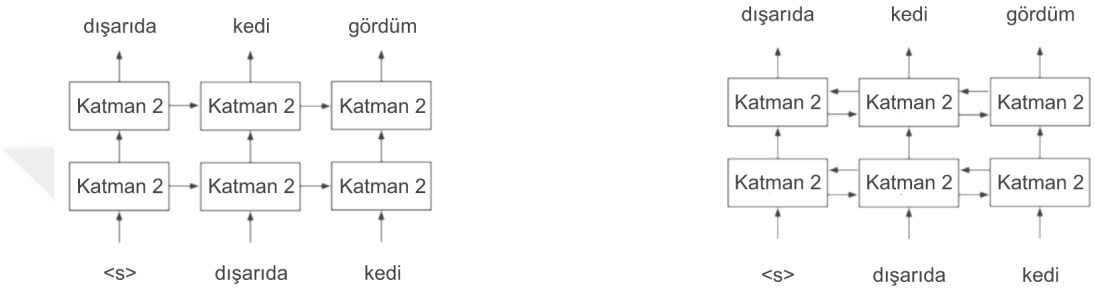
Dönüştürücü temel olarak dikkat mekanizmasını kullanan ve veriyi ağırlıklandırarak öğrenen bir derin öğrenme modelidir. Doğal dil işlemede ve bilgisayarda görü konularında sıkça kullanılmaktadır. Tekrarlayan sinir ağları gibi, dönüştürücüler sıralı girdileri işlemek için tasarlanmıştır. Fakat girdileri sırayla işlemesi gerekmez. Dikkat mekanizması yalnızca bağlam sağlamaktadır. Bu sayede girdiler paralel bir şekilde işlenmektedir ve eğitim süresi azalmaktadır.

Dönüştürücü dil modelleri ilk olarak Vaswani ve diğ. [81] tarafından makine çevirisi için tanıtılmıştır. Büyük miktarda etiketlenmemiş veri içeren büyük bir ağda dil modellerini önceden eğitme ve aşağı akış görevlerine hassas ayar yapma yöntemi, OpenAI GPT [62] ve BERT [17] gibi çeşitli doğal dil anlama görevlerinde çığır açmıştır. Önceden eğitilmiş BERT modelini kullanmak için orijinal çıktı katmanını göreve özel yeni bir katmanla değiştirmek ve tüm modelde hassas ayar yapmak gerekmektedir.

BERT dışındaki dil modelleri tek yönlü bağlamlardan oluştuğu için metinleri sola ya da sağa doğru işlemektedir. Aslında metin işlemenin yönlü olmaması veya çift yönlü

olması dil modeli açısında daha faydalı olacaktır. Çünkü bir kelime cümlede yalnızca sol tarafla ilgili veya yalnızca sağ tarafla ilgili değildir. Her iki tarafla ilgili olduğu için çift yönlü olması gerekmektedir. BERT, bu sebeple dönüştürücü kodlayıcı kullanmaktadır. Çift yönlü kodlayıcı kullanmaktadır.

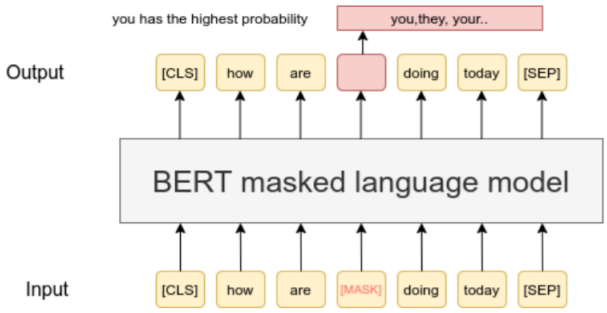
Tek yönlü modeller, temsilleri kademeli olarak öğrenmektedir. Türkçe metnin soldan sağa doğru okunduğu gibi, tek yönlü modeller de kelimeleri sırasıyla okuyarak öğrenmektedir. Çift yönlü modeller, hedef kelimeyi hem soldan bağlamdan, hem sağ bağlamdan öğrenmektedir. Çift yönlü modellerde Şekil 2.2'de görüldüğü üzere kelimeler kendilerini görebilmektedir. İleri yönlü ve geri yönlü iki adet tekrarlayan sinir ağı bulunmaktadır.



Şekil 2.2: Tek yönlü ve çift yönlü bağlam.

BERT çalışmasında temel olarak iki görev bulunmaktadır. Bunlardan bir tanesi maskeli dil modeli, diğeri ise sonraki cümle tahminidir.

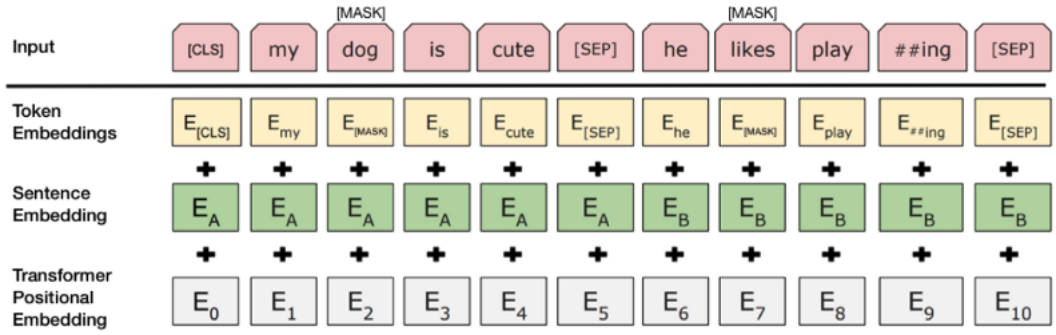
Maskeli dil modeli, metindeki kelimelerin belirli bir yüzdesini maskelemek ve bu kelimeleri tahmin etmeye çalışmaktır. Farklı yüzdelerle yapılabilir. BERT çalışmasında bu oran %15 olarak belirlenmiştir. Çalışmada çok az maskeleme yapıldığı zaman eğitimin çok pahalı ve uzun sürdüğü görülmüştür. Çok fazla maskeleme ise bağlamları kaybetmeye sebep olmuştur. Bu sebeple optimum seviye %15 olarak seçilmiştir. Şekil 2.3'de örnek olarak bir cümlede maskelenmiş bir kelime ve o kelimenin yerine gelme olasılığı olan kelimeler gösterilmiştir.



Şekil 2.3: BERT maskeli dil modeli [37].

Maskeleme yapıldığı zaman, maskelenen kelime hiçbir zaman hassas ayar aşamasında görülmediği için probleme yol açmaktadır. Buna çözüm olarak maskelenen kelime oranı sabit tutulmuştur fakat tüm maskeleri [MASK] etiketi ile değiştirmemişlerdir. Maskelerin %80'i [MASK] etiketi olarak koyulmuştur. %10'u farklı rastgele bir kelime ile değiştirilmiştir. %10'u da aynı şekilde bırakılmıştır. Bu sayede maskelenen kelime artık hassas ayar sırasında görülmektedir.

BERT çalışmasında ikinci görev olan sonraki cümle tahmini, cümleler arasındaki ilişkileri öğrenmek için B cümlesinin A cümlesinden devam eden gerçek cümle mi yoksa rastgele bir cümle mi olduğunu tahmin etmeye çalışmaktadır. Yani iki cümle arasında bir bağ olup olmadığını öğrenen yöntemdir. Bu görev için girdi olarak 30.000 kelime kullanılmıştır. Şekil 2.4'de görüldüğü üzere her kelime pozisyon vektörü, segment vektörü ve belirteç vektörünün toplamından oluşmaktadır.

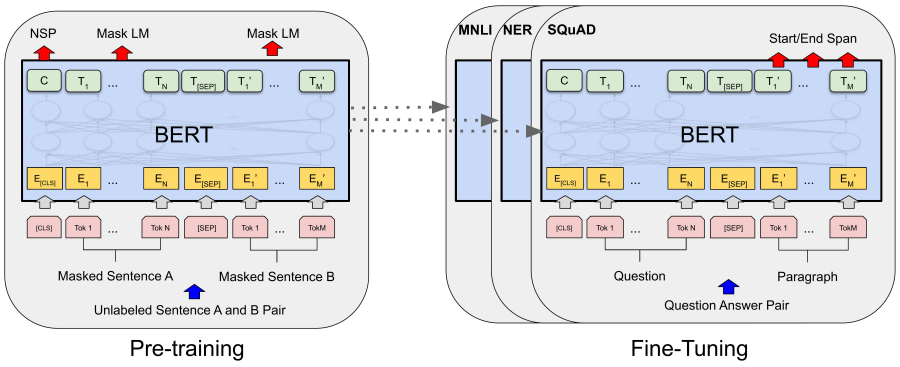


Şekil 2.4: Sonraki cümle tahmini [17].

Dikkat mekanizması, makine çevirisi görevinin performansını arttırmak amacıyla geliştirilmiştir. Amacı, kod çözücünün en uygun vektörlere en yüksek ağırlıkları vererek, kodlanan tüm girdi vektörlerinin ağırlıklı bir kombinasyonu ile esnek bir şekilde giriş dizisinin en ilgili kısımlarını kullanmasını sağlamaktır.

Önceden eğitilmiş dönüştürücü dil modelini kullanmanın en yaygın yolu, Şekil 2.5'de görüldüğü üzere orijinal çıktı katmanını göreve özel yeni bir katmanla değiştirmek ve tüm modeli hassas ayar yapmaktır. Bu aşama ile yeni çıktı katmanı parametreleri öğrenilir ve tüm orijinal ağırlıklar değiştirilir. Ayrıca sözcük temsil ağırlıkları ve dönüştürücü blokları da değişir. Örneğin, metin sınıflandırma görevinde, eklenen bir doğrusal sınıflandırıcı, [CLS] gömmeyi çıktı sınıfları üzerinde normalleştirilmemiş bir olasılık vektörüne yansıtır. Bu süreç, yeni çıktı katmanının ağırlık başlatması ve stokastik hassas ayar optimizasyonunda veri sırası olarak iki rastgelelik kaynağı sunmaktadır.

Hassas ayar, özellikle 10.000'den daha küçük veri setlerini kullanarak yapıldığında önemli derecede başarılı sonuçlar vermektedir. Dil modelini bu şekilde kullananlar, birçok hassas ayar denemesi yapmaktadır ve doğrulama performansına göre en iyi modeli seçmektedir.



Şekil 2.5: BERT hassas ayar [17].

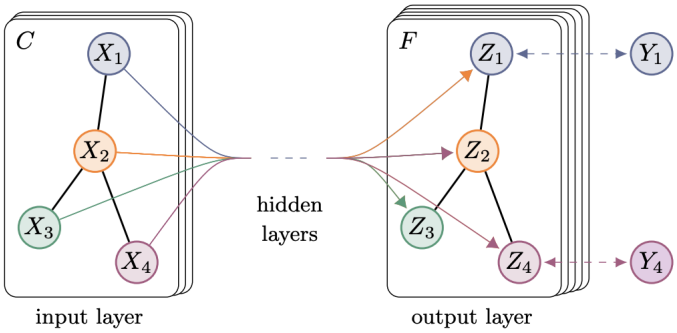
2.3 Çok Katmanlı Algılayıcı (MLP [23])

Çok katmanlı algılayıcı (MultiLayer Perceptron MLP), bir ileri beslemeli yapay sinir ağı sınıfıdır. MLP terimi belirsiz bir şekilde, bazen gevşek bir şekilde herhangi bir ileri beslemeli anlamına gelir, bazen katı bir şekilde birden çok algılayıcı katmanından (eşik aktivasyonlu) oluşan ağları ifade etmek için kullanılır. Çok katmanlı algılayıcılar, özellikle tek bir gizli katmana sahip olduklarında, bazen halk dilinde "vanilya" sinir ağları olarak adlandırılır.

Bir MLP, en az üç düğüm katmanından oluşur: bir giriş katmanı, bir gizli katman ve bir çıkış katmanı. Giriş düğümleri dışında, her düğüm, doğrusal olmayan bir aktivasyon işlevi kullanan bir nörondur. MLP, eğitim için geri yayılım adı verilen denetimli bir öğrenme tekniği kullanır. Çoklu katmanları ve doğrusal olmayan aktivasyonu, MLP'yi doğrusal bir algılayıcıdan ayırır. Doğrusal olarak ayrılamayan verileri ayırt edebilir. Bu çalışmada MLP algoritması metin verisi ile konum tespiti yapılması ve BERT ile karşılaştırılması amacıyla kullanılmıştır.

2.4 Evrişimsel Çizge Ağı (GCN [40])

Evrişimsel Çizge Ağı (Graph Convolutional Networks GCN), grafik yapılı veriler üzerinde yarı denetimli öğrenme için bir yaklaşımdır. Doğrudan grafikler üzerinde çalışan evrişimli sinir ağlarının verimli bir çeşidine dayanmaktadır. Evrişimsel mimarinin seçimi, spektral grafik evrişimlerinin yerelleştirilmiş birinci dereceden yaklaşımıyla motive edilir. Şekil 2.6'de görüldüğü üzere model, grafik kenarlarının sayısında doğrusal olarak ölçeklenir ve hem yerel grafik yapısını hem de düğümlerin özelliklerini kodlayan gizli katman temsillerini öğrenir. GCN modeli bu araştırmada konum tespitinde temel alınması ve BERT ile karşılaştırılması amacıyla kullanılmıştır.



Şekil 2.6: Evrişimsel çizge ağı (GCN) [40].

2.5 Twitter Verileri

Bu çalışmada veri mühendisliği amacıyla incelenen tüm görevlerde metin verisi kullanılmaktadır. Sosyal medya metinleri ise tüm araştırmaların odak noktası olmuştur. Bu sebeple Twitter üzerinden veri toplanmış ve kullanılmıştır. Twitter kullanıcıları her gün milyonlarca tweet atmaktadır. Twitter tarafından araştırmacılara, kullanıcı bilgisi, kullanıcı ağı ve paylaştığı metinler gibi içerikler sınırlı olarak sunulmaktadır. Fakat sınırlı veri kullanmak araştırmaların zayıf noktası olmuştur. Bu sebeple araştırmacılar kendi çabaları ile veri toplama araçları yazmaktadır veya açık kaynaklı veri toplama araçları kullanmaktadır.

Tweet, Twitter'da paylaşılan bir gönderidir. Bir tweet en fazla 280 karakter uzunluğunda olabilir. Metin, URL ve etiketler ve bahsetmeler içerebilir. Video ve fotoğraf barındırabilir. Twitter API tarafından sunulan her bir tweet'in içinde, id, tarih, saat, tarih, saat dilimi, kullanıcı id, kullanıcı adı, yer, tweet, bahsetmeler, linkler, fotoğraflar, yanıt sayısı, beğeni sayısı gibi 30'a yakın bilgi bulunmaktadır. Twitter'ın sunduğu kullanıcı bilgisinde ise kullanıcı adı, biyografi, lokasyon, link, katılma tarihi, tweetleri, takipçileri ve takip ettikleri gibi 20'ye yakın bilgi bulunmaktadır. Şekil 2.7'de anonimleştirilmiş bir tweet bilgisi bulunmaktadır.

Bu çalışmada kullanılan açık kaynaklı tweet toplama aracı olan Twint, tüm bu bilgileri kullanıcı adıyla veya kelime sorgusuyla veri taraması yaparak kaydetmektedir. Veri toplama sınırı bulunmadığı için aranan kullanıcılar veya sorgular için yüksek miktarda veri kaynağı sunmaktadır.

```
{
    "id":1442925253978570760,
    "conversation_id":1441862573247262720,
    "created_at":"2021-09-28 21:52:43 +03",
    "date":"2021-09-28",
    "time":"21:52:43",
    "timezone":300,
    "user_id":1213481702,
    "username":"ahmet2121",
    "name":"ahmet çelebi",
    "place":"Ankara",
    "tweet":"Vay be, ne günler geçiriyoruz #pazar",
    "language":"tr",
    "mentions":"[]",
    "urls":"[]",
    "photos":"['https://pbs.twimg.com/media/FAZM3VrVgAxaaaT.jpg']",
    "replies_count":2,
    "retweets_count":3,
    "likes_count":12,
    "hashtags":"[pazar]",
    "cashtags":"[]",
    "link":"https://twitter.com/ahmet2121/status/1442925253777888999",
    "retweet":false,
    "quote_url":"0",
    "video":1,
    "thumbnail":"https://pbs.twimg.com/media/aaaM3VrVgAAt4yT.jpg",
    "near":null,
    "geo":null,
    "source":null,
    "user_rt_id":null,
    "user_rt":null,
    "retweet_id":null,
    "reply_to":"[{'screen_name': 'mehmet21', 'name': 'memed', 'id': '1305059434461376999'}]",
    "retweet_date":null,
    "translate":null,
    "trans_src":null,
    "trans_dest":null
}
```


Şekil 2.7: Örnek tweet verisi.

3. İLGİLİ ÇALIŞMALAR

Bu bölümde bu araştırmayla ilgili çalışmalar konu bazında incelenecektir. Öncelikle 3.1 bölümünde dönüştürücü dil modelleri ve 3.2 bölümünde hassas ayar çalışmaları anlatılmaktadır. Ardından 3.3 bölümünde eğitim verisinin modele etkisini inceleyen çalışmalar, 3.4 bölümünde kontrole değer iddiaların tespiti çalışmaları, 3.5 bölümünde taraf tespit çalışmaları ve son olarak 3.6 bölümünde konum tespit çalışmaları incelenecektir.

3.1 Dönüştürücü Dil Modelleri

Araştırmacılar dönüştürücü dil modellerini, göreve uyarlanabilir ön eğitim [26, 34, 61, 77] ve domaine uyarlanabilir ön eğitim [26, 45] olarak iki farklı şekilde eğitmeye devam etmiştir.

3.2 Dönüştürücü Dil Modellerine Hassas Ayar

Önceden eğitilmiş modeller yaygınlaştıktan sonra, bu modellere nasıl daha iyi hassas ayar yapılabileceğini araştıran araştırmalar olmuştur [8, 19, 34, 63, 91, 92]. Zaken ve diğ. [8], küçük-orta eğitim verileriyle, yalnızca önceden eğitilmiş BERT modellerinin önyargı terimlerinde hassas ayar yapmanın tüm modelde hassas ayar yapmakla rekabet edebileceğini göstermiştir. Eisenschlos ve diğ. [19], uygulayıcıların dil modellerini kendi dillerinde verimli bir şekilde eğitmelerini ve hassas ayar yapmalarını sağlamak için çok dilli dil modeli hassas ayarı önermiştir. Ek olarak, önceden eğitilmiş mevcut bir çapraz dilli modeli kullanarak sıfır atış yöntemi önermiştir. Howard ve diğ. [34], doğal dil işlemedeki herhangi bir göreve uygulanabilen etkili bir aktarım öğrenme yöntemi olan Evrensel Dil Modeli İnce Ayarını önerir ve bir dil modelinde hassas ayar yapmak için anahtar olan ayrımcı hassas ayar, eğik üçgen öğrenme oranları ve kademeli olarak çözme gibi teknikleri sunmuştur. Radya et diğ. [63], hassas ayarlanmış modellerin parametre uzayında önceden eğitilmiş modele yakın olduğunu ve yakınlığın katmandan katmana değiştiğini sunmuştur. Devasa dil modellerinde hassas ayarın, önceden eğitilmiş parametrelerin belirli katmanlarındaki belirli sayıda girişi basitçe sıfıra ayarlayarak, hem göreve özel parametre depolama hem de hesaplama maliyetinden tasarruf sağlanabileceğini göstermiştir. Zengin ve diğ. [91] eğitim setini değiştirerek dönüştürücü dil modellerinde etkili bir şekilde hassas ayar yapmayı göstermiştir. Araştırdıkları yöntemler arasında dile özgü eğitim, zayıf denetim, makine çevirisiyle veri büyütme, alt örnekleme ve diller arası eğitim bulunmaktadır.

Zhang ve diğ. [92], birkaç örnek senaryoda yaygın olarak gözlenen kararsızlıklara odaklanarak, BERT bağlamsal temsillerinin hassas ayarına ilişkin bir çalışma yürütmüştür. Çalışmada kararsızlığa neden olan faktörlerin, yanlı gradyan tahmini ile standart olmayan bir optimizasyon yönteminin yaygın kullanımı, aşağı akış görevleri için BERT ağının önemli bölümlerinin sınırlı uygulanabilirliği, önceden belirlenmiş ve az sayıda eğitim yinelemesi kullanmanın yaygın uygulaması olarak belirtilmiştir.

3.3 Eğitim Verisinin Etkisini İnceleyen Çalışmalar

Bu bölümde eğitim verisinin önemi üzerine yapılan çalışmalar 5 farklı alt başlıkta anlatılacaktır. Eğitim verisini değiştirmek için çapraz dil çalışmaları, zayıf denetim çalışmaları, geri çeviri çalışmaları, aşırı örnekleme çalışmaları ve aktif öğrenme çalışmaları incelenmiştir.

3.3.1 Çapraz dil çalışmaları

Dönüştürücü dil modelleri doğal dil işlemede büyük başarı kaydetmiştir. BERT [17] ve RoBERTa [46] gibi tek dilli önceden eğitilmiş dil modellerinin başarısı, çalışmaları çok dilli dönüştürücü dil modellerine sevk etmiştir. mBERT [17], XLM [44], XLM-R [13] gibi çok dilli önceden eğitilmiş modeller, çapraz dilli görevlerde en iyi sonuçları vermektedir.

Önceden eğitilmiş çok dilli dönüştürücü dil modelleri kullanılarak, doğal dil işleme görevlerini çapraz dilli bir şekilde çözmeyi öneren çalışmalar bulunmaktadır. Zheng ve diğ. [93] XTREME kıyaslaması üzerinde metin sınıflandırma, soru yanıtlama, sıra etiketleme gibi görevlerde çapraz dil çalışmalarının önemli ölçüde etkili olduğunu göstermiştir. Yu ve diğ. [89] sıfır atışlı (zero shot) ve birkaç atışlı (few shot) çapraz dil görevlerinde çok dilli önceden eğitilmiş dil modeli kullanmanın iyi sonuçlar verdiğini göstermiştir. Ayrıca hedef dildeki etiketlenmemiş verileri kullanarak çok dilli dil modellemeye tamamlayıcı olarak daha genelleştirilmiş anlamsal denklikleri öğrenmeyi amaçlayan ortak eğitime dayalı yeni bir hassas ayar yöntemi önermiştir.

3.3.2 Zayıf denetim çalışmaları

Eğitim verisini arttırmak model başarısını olumlu yönde etkilemektedir. Eğitim verisini manuel olarak etiketlemek maliyetli bir süreç olduğu için zayıf deneyim kullanılarak eğitim verisi arttırılmaktadır. Snorkel [66] eğitim verisi arttırma yöntemi olarak etiketleme fonksiyonu yazmayı önermektedir. Bu sayede kısa sürede büyük hacimli eğitim verileri oluşturmayı sağlamaktadır. Snuba [79] küçük bir eğitim verisi alarak sezgisel yöntemler kullanarak eğitim verisini otomatik olarak zayıf deneyimle arttırmayı sunmuştur.

Zayıf denetimle arttırılmış veriyi model eğitimi yerine, önceden eğitilmiş dönüştürücü dil modellerine hassas ayar yapmak için kullanan çalışmalar da [51, 90, 91] bulunmaktadır.

3.3.3 Veri büyütme çalışmaları

Veri büyütme, yapay zeka çalışmalarında kısıtlı veri bulunduğu zaman veriyi arttırmak amacıyla yapılan yöntemdir. Metin verisini anlamını bozmayacak şekilde karakter tabanlı, kelime tabanlı ve cümle tabanlı değişiklik yaparak arttırmak mümkündür [21]. Metin silme, ekleme, değiştirme gibi yöntemler kullanılabilir. Bununla birlikte metin arttırmak için geri çeviri ve aşırı örnekleme çalışmaları da bulunmaktadır.

Geri çeviri, hedef dilin kaynak dile çevrilmesi ve bir modeli eğitmek için hem orijinal kaynak cümlelerin hem de geri çevrilmiş cümlelerin karıştırılmasıdır. Böylece kaynak dilden hedef dile aktarılan eğitim verilerinin sayısı artırılabilir. Etiketleme maliyetli ve zaman alıcı bir süreçtir. Bu nedenle, etiketlenmiş veri boyutunu artırmak için geri çeviri yöntemi kullanılmaktadır. Geri çeviri yöntemi, metin sınıflandırma [75], makine çevirisi [95], metin üretme [74] gibi doğal dil işlemenin birçok görevinde kullanılmaktadır.

Aşırı örnekleme, veri dağılımında azınlık sınıfa ait verileri rastgele olarak çoğaltma işlemine denir. Eğitim verisinin dengesiz olması modelin öğrenme sürecini olumsuz etkilemektedir. Veri kümesini dengeli hale getirmek için oran olarak fazla olan sınıfın verisini azaltarak yetersiz örnekleme yapılabilir. Fakat bu etiketli veri miktarını azaltacağı için her zaman tercih edilmez. Bunun yerine oran olarak düşük olan sınıfa aşırı örnekleme yapmak, eğitim verisini dengeli hale getirecektir. Suh ve diğ. [76] aşırı örnekleme yöntemlerinin genellikle sınıflandırıcıların performansını iyileştirdiğini söylemektedir. Moreo ve diğ. [55] dağılım hipotezinin geçerli olduğu metin gibi verileri sınıflandırmak için özel olarak tasarlanmış yeni bir dağılımsal rastgele aşırı örnekleme yöntemi sunmuştur.

3.3.4 Aktif öğrenme çalışmaları

Aktif öğrenme, mümkün olan en az örneğe açıklama eklerken bir modelin performans kazancını en üst düzeye çıkarmaya çalışmaktadır. Derin öğrenme veriler için açgözlüdür ve model yüksek kaliteli özelliklerin nasıl çıkarılacağını öğrenecekse, çok sayıda parametreyi optimize etmek için büyük miktarda veri kaynağı gerektirir [67]. Doğal dil işleme görevlerinde genellikle etiketlenmemiş veri bulmak kolaydır fakat etiketleme süreci uzun ve maliyetli olmaktadır. Bu durumlarda, öğrenme algoritmaları aktif olarak kullanıcıya soru sorarak etiketleme yapabilir. Bu tür yinelemeli denetimli öğrenmeye aktif öğrenme denir. Aktif öğrenme sürecinde etikelenen veri miktarı, normal denetimli öğrenmeye göre oldukça düşük miktarda olmaktadır. Bu sayede daha az maliyetli ve daha kısa sürece, daha büyük veri oluşturulmaktadır. Tong ve diğ. [78] metin sınıflandırma çalışmasını destek vektör makineleri ile yaparken aktif öğrenme yöntemi kullanmıştır. Yang ve diğ. [87] ise yine metin sınıflandırma görevi için aktif öğrenme kullanarak efektif çoklu etiketleme yapmıştır. Schröder ve diğ. [72] derin sinir ağları kullanarak aktif öğrenme sürecini araştırmıştır.

3.4 Kontrole Değer İddiaların Tespiti Çalışmaları

İlk kontrole değer iddiaların tespit çalışmalarından biri ClaimBuster [33]'dır. Konuşma bölümü (POS) etiketleri, adlandırılmış varlıklar, duyarlılık ve iddiaların TF-IDF temsilleri gibi birçok özelliği kullanmaktadır. Patwari ve diğ. [59] öznitelik olarak 1976 ve 2016 arasındaki başkanlık tartışmalarını, POS etiketlerini, varlık geçmişini ve kelime hazinesini kullanmıştır. Gencheva ve diğ. [24], uzun bir cümle düzeyi listesi ve duygu, adlandırılmış varlıklar, sözcük yerleştirmeleri, konular, çelişkiler ve diğerlerini içeren bağlamsal özellikler içeren bir sinir ağı modeli önermiştir. Jaradat ve diğ. [38] Arapça için benzer özellikleri kullanarak Gencheva ve diğ. modelini genişletmiştir. Vasileva ve diğ. [80], bir iddianın saygın doğrulama kuruluşları tarafından doğrulanıp doğrulanmadığını tespit etmek için çok görevli bir öğrenme modeli önermiştir.

CLEF konferansı 2018 yılından itibaren CheckThat Lab (CTL) organize etmektedir. İngilizce ve Arapça dillerini kapsayan ilk organizasyon CTL'18'e yedi ekip katılmıştır. Ekipler, tekrarlayan sinir ağı (RNN) [29], çok katmanlı algılayıcı [96], rastgele orman (RF) [1], k en yakın komşu (kNN) [25] ve gradyan artırma [88] kelime çantası [96], karakter n-gram [25], POS etiketleri [29, 88, 96], sözlü formlar [96], adlandırılmış varlıklar [88, 96], sözdizimsel bağımlılıklar [29, 96] ve kelime yerleştirmeleri [29, 88, 96] gibi çeşitli modelleri kullanmışlardır. Prise de Fer ekibi [96], İngilizce veri kümesinde çok katmanlı algılayıcı öğrenme-SVM ile kelime çantası, POS etiketleri, adlandırılmış varlıklar, sözlü formlar, olumsuzlamalar, duyarlılık, yan tümceler, sözdizimsel bağımlılık ve kelime yerleştirmelerini kullanarak en iyi ortalama kesinlik (mean average precision MAP) puanlarını elde etmiştir. Arapça veri kümesinde, BigIR ekibi [88], özellik olarak POS etiketlerini, adlandırılmış varlıkları, duyguları, konuları ve kelime yerleştirmelerini kullanarak diğerlerinden daha iyi performans göstermiştir.

CTL'19'da sadece İngilizce için düzenlenen kontrol değerlilik görevine 11 ekip katılmıştır. Görevin katılımcıları, cümlelerin okunabilirliği ve bağlamları dahil olmak üzere birçok özelliğe sahip LSTM, SVM, naive bayes ve lojistik regresyon (LR) gibi birçok öğrenme modelini kullanmıştır [5]. Kopenhag ekibi [30], zayıf denetimli LSTM modeliyle sözdizimsel bağımlılık ve sözcük yerleştirmelerini kullanarak en iyi genel ortalama kesinlik (MAP) puanını elde etmiştir.

CTL'20'de, kontrol edilebilirlik [6] için Görev 1 ve Görev 5 olmak üzere iki görev düzenlenmiştir. Görev 1 Arapça ve İngilizce tweetleri kapsarken, Görev 5 İngilizce tartışmaları kapsar. Görev 1'in katılımcıları BERT [2, 11, 32, 39, 82], RoBERTa [58, 82] BiLSTM [36, 49], CNN [2], RF [52], LR [39] ve SVM [11] gibi çeşitli özelliklere sahip modeller ve FastText [39, 52], Glove [49], PCA [11], TF-IDF [52], POS etiketleri [11, 39] ve adlandırılmış varlıklar [11] gibi öznitelikler kullanmışlardır. Accenture ekibi [82], Arapça için BERT modelini ve İngilizce için RoBERTa modelini kullanarak her iki veri kümesinde de en iyi ortalama kesinlik (MAP) puanını elde etti. Görev 5'in katılımcıları, TF-IDF'li BERT [39], LR ve LSTM modellerini, kelime yerleştirme ve POS etiketi özelliklerini [7] kullandı. NLPIR01 Takımı, kelime yerleştirmeli LSTM modeli kullanılarak ilk sırada yer almaktadır. Ayrıca farklı örnekleme yöntemlerini araştırmışlardır ancak performansı iyileştirmediklerini bildirmişlerdir.

CTL'21'de lab dördüncü defa düzenlenmiştir. İngilizce, Türkçe, Arapça, İspanyolca ve Bulgarca olmak üzere 5 dil bulunmaktadır. 3 farklı görev tanımlanmıştır, görev 1

kontrole değer iddiaların tespitinden oluşmaktadır ve 15 takım katılmıştır [57]. Katılımcının en yüksek olduğu diller İngilizce ve Arapça olmuştur. Dört ekip tüm diller için laba katılmıştır. Arapça dilinde AraBERT gibi önceden eğitilmiş modellere hassas ayar yapılmıştır. Accenture ekibi [83], olumlu örneklerin sayısını artırmak için bir etiket büyütme yaklaşımı kullanmıştır. Bulgarca dilinde UPV ekibi [71] SBERT kullanarak başarılı sonuç almıştır. İngilizce'de en iyi sonucu alan ekip NLPIR@UNED [50], birkaç farklı dil modeline hassas ayar yapmıştır ve BERTweet'in en iyi sonuç verdiğini bildirmiştir. İkinci en iyi puan da aynı modeli kullanmıştır. İspanyolca'da TOBB ETU ekibi [91] makine çevirisi ve zayıf denetim dahil olmak üzere farklı veri büyütme stratejilerini incelemiştir fakat en iyi sonucun veri arttırmadan kullanılan BETO olduğunu bildirmiştir. Türkçe için en iyi sonucu alan TOBB ETU ekibi [91], kullanıcı bahsetmeleri ve URL'leri kaldırdıktan sonra BERTurk'e hassas ayar yaptığı bildirmiştir.

3.5 Taraf Tespit Çalışmaları

Geçmiş çalışmalar incelendiğinde iki tür taraf tespit çalışması gözlemlenmektedir. Hasan ve diğ. [31], Mohammad ve diğ. [53] ve Ebrahimi ve diğ. [18] çalışmalarında tek hedefe özel sınıflandırma yapmıştır. Öte yandan, Xu ve diğ. [86], çok hedefli taraf tespiti üzerine çalışmaktadır. Taraf tespit sürecinin başarısını artırmak için, taraf tespiti yalnızca varlık ismi tanıma [42] ile değil, aynı zamanda çoğunluk oyu sınıflandırıcıları [12] ile de yapılmıştır. Bu yaklaşımı kullanmak için, farklı konularda kaliteli veri gerekmektedir. Her konu için ayrı etiketleme ve model üretilmesi gerekmektedir. Bu nedenle konuya özel yaklaşım yerine Xu ve diğ. [86] gibi hedefler arası sınıflandırma çalışmaları yapılmıştır. Buradaki amaç, modeli genelleştirmek ve kaynak hedef konu dışında bir hedef konu hakkında farklı bir taraf öğrenmektir. Çapraz hedeflerde taraf tespiti, yeni bir konu için eğitim verisi gerektirmez. Ancak modelin eğitiminde kullanılan konu ile yeni konu arasındaki ilişki olmalıdır. Her iki konu da benzer veya ilgili bir alanda olmalıdır. Hedefler arası çözümlerin kullanılmasının belirli alanlarda taraf tespiti performansını iyileştirdiği gösterilmiştir, ancak nihai bir çözüm sunmaz.

Bu iki yaklaşım üzerinden sıfır atışlı öğrenme yöntemini kullanarak daha genel bir çözüm sunmayı amaçlayan çalışmalar da mevcuttur. Allaway ve McKeown [3] yaptıkları sıfır atışlı çalışmada İngilizce dilinde başarılı sonuçlar almıştır. Fakat bilindiği kadarıyla Türkçe dili için literatürde sıfır atışlı çalışma bulunmamaktadır.

Bunun yanında belirlenecek verilerin dilinin de değişebileceği yadsınamaz bir gerçektir. Bu nedenle Lai ve diğ. [43], sosyal medyada paylaşılan siyasi tartışmaları inceleyerek belirli bir taraf tespit sisteminin farklı dillerde taraf tespitini nasıl gerçekleştirdiğini araştırmış ve performanslarını karşılaştırmış. Belirli bir hedef üzerinde iki farklı dilde çeşitli makine öğrenme modelleri kullanılarak taraf tespiti performansının değişiklik gösterdiği gözlemlenmiştir.

Taraf tespit çalışmaları çeşitlendikçe veri setleri ile ilgili çalışmaların sayısı da artmıştır. Bu çalışmalar arasında eğitim verilerinde bulunamayan ancak test verilerinde bulunan hedefler ile hem eğitim hem de test verilerinde bulunan hedefler için yapılan taraf tespit çalışmaları bulunmaktadır. Bu iki yöntemin performanslarının karşılaştırılması üzerine yapılan önemli çalışmalardan birini Mohammad ve diğ. [53] yapmıştır. Bu çalışmada iki farklı deney önerilmiştir. Birincisi, 0.70 verinin eğitim verisi olarak

kullanıldığı standart denetimli bir sınıflandırma modeli, ikincisi ise tüm verilerin test olarak kullanıldığı ve eğitim verisi olmadığı bir görevdir. Sonuç olarak, eğitim verilerine henüz dahil edilmeyen hedeflere yönelik taraf tespit çalışmalarının geleneksel yöntemin başarısına göre yetersiz kaldığı gözlemlenmiştir.

Veri kümeleri alanında yapılan çalışmalara bir başka örnek de Darwish ve diğ. [16] çalışmasıdır. Bu çalışma, Twitter kullanıcılarının tartışmalı konulardaki tutumlarını belirlemek için boyut küçültme ve kümeleme kullanan denetimsiz bir çerçeveyi incelemektedir. Veri çarpıklığı durumlarında bu çerçevenin avantajlı olduğu belirtilmiştir. Bazı veriler ve duruşlar, özellikle veri kümesi etiketlemede kullanıcılara zaman kazandırdığı ve alana özel bilgi gerektirmediği için veri kümesinde büyük miktarda yer kapladığında oluşur.

Küçük ve diğ. [41], Türkçe metinler üzerine bir çalışma olması ve kullanılan veri setini paylaşması nedeniyle bu çalışmada temel alınan bir çalışma olmuştur. Çalışmada destekleyici, karşıt ve tarafsız olmak üzere üç sınıf bulunmaktadır. Türkçe tweetlerden oluşan bir veri seti üzerinden iki popüler spor kulübü hedef olarak seçilmiştir. Bu iki hedef için kesinlik başarısı 0,82 ve 0,73 olarak elde edilmiştir.

3.6 Konum Tespit Çalışmaları

Sosyal medya kullanıcılarının konumunu bulmayı hedefleyen göreve konum tespiti denmektedir. Bu çalışmalar temel olarak kullanıcının yaşadığı yerin tespiti veya paylaşım konumunun tespiti olarak ikiye ayrılabilir. Konum tespiti temel olarak dört bilgi kullanılarak yapılmaktadır. Bunlar, kullanıcının profil bilgisi, kullanıcının oluşturduğu içerik, kullanıcının sosyal ağı ve zamansal bilgisidir. Farklı kombinasyonların kullanıldığı hibrit çalışmalar da bulunmaktadır.

Rahimi ve diğ. [65] 3 farklı veri setinde hem metin tabanlı hem de sosyal ağ tabanlı konum tespiti çalışması yapmıştır. Eğitim verisini ve kodunu paylaşması ve en yüksek performanslardan birini sunması nedeniyle bu çalışmada temel olarak kullanılan bir yöntem olmuştur. Han ve diğ. [28] kullanıcı konum tespitini metin tabanlı tahminlerle bulmayı hedeflemiştir. Zheng ve diğ. [94] kullanıcı konum tespiti için hibrit bir mekanizma kullanmıştır. Huang ve Carley [35] 20 milyon kullanıcıdan toplanan 40 milyardan fazla tweete dayalı olarak Twitter'da coğrafi etiketleme davranışına ilişkin ampirik bir çalışma yürütmüştür. Bu çalışmaya göre farklı ülkelerin coğrafi etiket kullanım oranları karşılaştırılmıştır. Örneğin Kore'de %3 olan oran, Endonezya'da %40'lara çıkmaktadır. Lourentzou ve diğ. [47], sinir ağlarının coğrafi konum tespitine uygulanmasını incelemiştir ve yalnızca metne dayalı konum tespitinde sinir ağlarını geliştirmek için birden fazla teknik uygulamışlardır. Üç farklı Twitter veri kümesinde, uygun ağ mimarisi kullanmanın, aktivasyon fonksiyonunun ve batch normalleştirmenin bu görevdeki performansı iyileştirdiği raporlanmıştır. Flatow ve diğ. [22], coğrafi bilgisi bulunmayan metinleri içeriğine göre belirleme çalışması yapmıştır. Metinlerin n-gramlarını ve konum dağılımlarını modelleyerek yerel bir coğrafi alanla ilişkilendirmektedir.

4. VERİ TOPLAMA ve VERİYİ İŞLEME

Bu bölümde çalışmada kullanılan verilerin nasıl toplandığı ve ne şekilde işlendiği açıklanacaktır. Sırasıyla 4.1 bölümünde kontrole değer iddiaların tespiti çalışmasında, 4.2 bölümünde taraf tespiti çalışmasında ve 4.3 bölümünde konum tespiti çalışmasında kullanılan veriler incelenecektir.

4.1 Kontrole Değer İddiaların Tespiti Veri Seti

Kontrole değer iddiaların tespiti çalışması CLEF21 Check That [56] tarafından paylaşılan veri seti kullanıkarak yapılmıştır. Paylaşılan veriler İngilizce, Türkçe, İspanyolca, Bulgarca ve Arapça olmak üzere 5 farklı dil için sunulmuştur. Veriler eğitim, test ve geliştirme verisi olarak bölünmüş bir şekilde, her dil için kendine özel bir konuda verilmiştir.

4.1.1 Veri istatistiği

Veri istatistikleri ve veri dağılımı Çizelge 4.1'de bulunmaktadır. Arapça'da 4700, Bulgarca'da 3707, İngilizce'de 1312, İspanyolca'da 4990 ve Türkçe'de 3300 olmak üzere toplam 18009 veri kullanılmıştır.

Çizelge 4.1: Kontrole değer iddiaların tespiti veri istatistiği.

Dil	Konu	Eğitim		Geliştirme		Test	
		TM	TM Değil	TM	TM Değil	TM	TM Değil
Arapça	Karışık	763	2676	265	396	242	358
Bulgarca	Covid 19	392	2608	62	288	76	281
İngilizce	Covid 19	290	532	60	80	19	331
İspanyolca	Siyaset	200	2295	109	1138	120	1128
Türkçe	Karışık	729	1170	146	242	183	830

4.1.2 Zayıf denetim verisi

Etiketlenmiş veri boyutunu artırmanın bir başka yolu da zayıf denetimdir. Bu nedenle zayıf deneyim yöntemi amacıyla ekstra veri çekilmiş ve kullanılmıştır. Öncelikle her veri kümesindeki kelimeler sıklıklarına göre sıralanmıştır ve en sık kullanılan 100 kelime arasından ilgili veri kümelerinin konusuyla ilgili 10 kelime manuel olarak seçilmiştir. Şekil 4.1 bu anahtar kelimeleri göstermektedir. Ardından, Twint aracı kullanılarak bu kelimelerin her biri için ayrı ayrı 500 tweet çekilmiştir. Bunun sonucunda her dilden 5000 olmak üzere toplam 25000 ekstra veri çekilmiştir.

Arapça	Bulgarca	İngilizce	İspanyolca	Türkçе
كورونا	България	covid19	españa	yüzde
النسويات	случаи	virus	gobierno	milyar
بفيروس	заразени	people	millones	türkiye
النسوية	кризата	cases	sánchez	dolar
الشعب	пандемията	health	personas	istanbul
الصحة	ваксина	testing	euros	belediye
إصابة	разпространението	confirmed	gobierno	ticaret
عاجل	мерките	coronavirus	madrid	seçim
وزارة	европа	hospital	política	ülkeler
التطبيع	денонощие	patients	contra	enflasyon

Şekil 4.1: Veri çekme amacıyla seçilen kelimeler.

4.2 Taraf Tespiti Veri Seti

Taraf tespiti deneylerini yapabilmek için farklı konuları, dilleri ve hedefleri kapsayan çeşitli veri kümelerine ihtiyaç vardır. Fakat Türkçe taraf tespiti için mevcut veri setleri oldukça sınırlıdır. Küçük ve Can [42] tarafından Türkiye'deki iki popüler futbol kulübü Fenerbahçe ve Galatasaray için ikili etiketli (destekleyen ve karşıt) 1065 tweet içeren tek bir veri seti bulunmaktadır. Ayrıca Twitter'ın tweet içeriklerini yeniden dağıtma yasağı nedeniyle çalışma sadece tweet ID'leri paylaşılmıştır. Silinen tweetler ve kapatılan hesaplar nedeniyle sadece 454 tweet çekilmiştir. Bu nedenle, taraf tespiti için çeşitli konu ve hedefleri kapsayan yeni veri kümeleri oluşturulmuştur.

4.2.1 Veri etiketleme

Manuel etiketleme prosedürü için ikili etiketler (destekleyen ve karşıt) kullanılmıştır. Her tweet iki etiketleyici tarafından etiketlenmiştir. Etiketleyicilerin ortak karar almadığı metinler filtrelenmiştir, kullanılmamıştır. Ayrıca, birden fazla hedefi olan (örneğin, "FB'den nefret ediyorum ama GS'yi seviyorum") çelişkili metinler ve tarafsız bir duruş içeren metinler de kullanılmamıştır. Tüm etiketleyiciler, ana dili Türkçe olan ve aynı zamanda İngilizce bilen lisansüstü öğrencilerdir. Toplamda dört etiketleyici katkıda bulunmuştur.

4.2.2 Veri dağılımı ve veri istatistiği

Futbol, sağlık, ekonomi ve siyaset olmak üzere dört farklı alanda Türkçe veri seti oluşturulmuştur. Her veri kümesi için, farklı anahtar kelime kümeleriyle Twint kütüphanesini kullanarak tweetler çekilmiştir. Ardından, bir önceki bölümde açıklanan etiketleme prosedürü izlenerek etiketlendirilmiştir.

Futbolla ilgili tweet'lerin bulunduğu veri setinde, hedefleri belirli olan bir veri seti oluşturulmuştur. Daha fazla futbol kulübü için, daha fazla tweet manuel olarak etiketlenmiştir ve Küçük ve Can [42]'nin veri seti genişletilmiştir.

Fenerbahçe (FB), Galatasaray (GS), Beşiktaş (BJK) ve Trabzonspor (TS) olmak üzere dört takıma odaklanılmıştır ve her biri için 1000 tweet çekilmiştir. Anahtar kelime olarak takımların isimleri kullanılmıştır. Futbol veri setini oluşturmak için Küçük ve Can [42]'nin veri seti ile etiketlenen metinler birleştirilmiştir.

Sağlık, siyaset ve ekonomi alanındaki veri setleri için belirli bir hedef kullanılmamıştır, sadece metinlerde tek bir hedef olmasına ve destekleyen veya karşıt ifade olmasına dikkat edilmiştir. Çizelge 4.2, her veri kümesi için tweet çekmek için kullanılan anahtar kelimeleri ve her bir etiketin sayısını göstermektedir.

Zayıf denetim deneyinde otomatik etiketleme için şikeci ve şampiyon gibi destekleyici ve karşıt kelimelerin yanına kulüp isimleri eklenerek her takım için 10.000 tweet çekilmiştir. Bu verileri çekmek için Twint kütüphanesi kullanılmıştır.

Çizelge 4.2: Farklı domainlerin Türkçe verileri ve etiket dağılımı.

Konu	**Hedefe Özel**	**Anahtar Kelimeler**	**Destekleyici**	**Karşıt**	**Toplam**
Futbol	Evet	beşiktaş, galatasaray, trabzonspor, fenerbahçe	416	414	830
Sağlık	Hayır	covid, aşı, maske, fahrettin koca	475	765	1240
Ekonomi	Hayır	bitcoin, kripto para, dolar, ekonomi	620	1215	1835
Siyaset	Hayır	akp, mhp, chp	276	300	576

Çizelge 4.3, Futbol veri seti için kullanılan veri sayısını ve dağılımını içermektedir. Bu tabloda her takımın destekleyici ve karşıt sayıları ayrıntılı olarak verilmiştir. Ayrıca Küçük ve Can [42] çalışmasından çekilen verilerin dağılımı da bulunmaktadır.

Çizelge 4.3: Futbol kulüplerinin veri ve etiket dağılımı.

Hedef	**Kaynak**	**Destekleyici**	**Karşıt**	**Toplam**
Galatasaray	Manuel	86	20	106
Fenerbahçe	Manuel	27	42	69
Beşiktaş	Manuel	52	48	100
Trabzonspor	Manuel	37	64	101
Galatasaray	Küçük ve Can[42]	115	132	247
Fenerbahçe	Küçük ve Can[42]	99	108	207

Çizelge 4.4, İngilizce veri seti dağılımlarını içermektedir. Sağlık ve ekonomi veri seti sadece diller arası deneylerde kullanıldığından manuel etiketleme sayısı düşük tutulmuştur. Politika, feminizm, iklim değişikliği ve kürtaj veri setleri SemEval [54] çalışmasından alınmıştır.

Çizelge 4.4: İngilizce veri ve etiket dağılımı.

Konu	Kaynak	Destekleyici	Karşıt	Toplam
Sağlık	Manuel	20	15	35
Ekonomi	Manuel	52	13	65
Siyaset	SemEval [54]	305	832	1137
Feminizm	SemEval [54]	58	183	241
İklim Değişikliği	SemEval [54]	123	11	134
Kürtaj	SemEval [54]	46	189	235
Ateizm	SemEval [54]	124	464	588

4.3 Konum Tespiti Veri Seti

Bu bölümde konum tespiti için kullanılan veri açıklanacaktır. Bölüm 4.3.1'de veri çekme, Bölüm 4.3.2'de veri ön işleme, Bölüm 4.3.3'de veri istatistiği bulunmaktadır.

4.3.1 Veri çekme

Şehir bazlı sosyal medya verileri Twitter üzerinden toplanmıştır. Twitter'da konum bazlı veri çekmek için iki farklı yöntem izlenebilir. İlk yöntem, eğer kullanıcı atılan tweet için konum paylaşımını açtıysa verideki bu alan kullanılarak konum bazlı veri çekilebilmektedir. Fakat bu şekilde bulunan tweet sayısı oldukça az olduğu için yeterli veri toplanamaz. Bu sebeple ikinci yöntem uygulanmıştır. İkinci yöntem, eğer kullanıcı profiline bir konum bilgisi girdiyse, bu bilgi kullanılarak kullanıcı adları ve tweet'leri toplanabilir. Eğer kullanıcı bir ilçe adı girerse, bulunduğu ilin ismi kullanıldığında bu kullanıcı da gelmektedir. Ayrıca üçüncü bir yöntem olarak koordinat kullanılarak ve bir daire yarıçapı uzunluğu verilerek de veri çekilmektedir.

Veri çekmek için açık kaynaklı bir Twitter veri kütüphanesi olan Twint kullanılmıştır. Bu kütüphane ile öncelikle 81 il için her şehirden 5000 tweet olmak üzere toplam 405.000 tweet çekilmiştir. Bu tweetler 81 satırdan oluşan bir bash script ile çekilmiştir. Ankara şehri için örnek bash script: "twint –near "Ankara" –limit 5000 -o Ankara.csv –csv" şeklindedir.

Ardından bu tweetleri paylaşan kullanıcı listesi oluşturulmuştur. Toplam 104.143 kullanıcı bulunmuştur. Her şehir için ortalama kullanıcı sayısı 1285 olmuştur.

Ardından her şehir için kullanıcı sayısı 500 olarak filtrelenmiştir. Bu 500 kullanıcının son paylaştığı 500 tweet çekilmiştir. Her şehir için 500 satırlık bir bash script çalıştırılmıştır. Bir kullanıcının 500 tweetini çeken bash script: "twint -u username –timeline –limit 500 –output username.csv –csv" şeklindedir.

Bunun sonucunda toplam 40.500 kullanıcı ve 16.116.035 tweet elde edilmiştir. Her kullanıcının 500 tweeti olmadığı için hesaplanandan düşük bir sayı çıkmıştır. Bu tweetlerin içerisinde tweetin atıldığı koordinat bilgisini tutan yalnızca 64.044 tweet ve 1357 kullanıcı bulunmaktadır. Koordinat bilgisi yeteri kadar olmadığı için bu veri kullanılmamıştır.

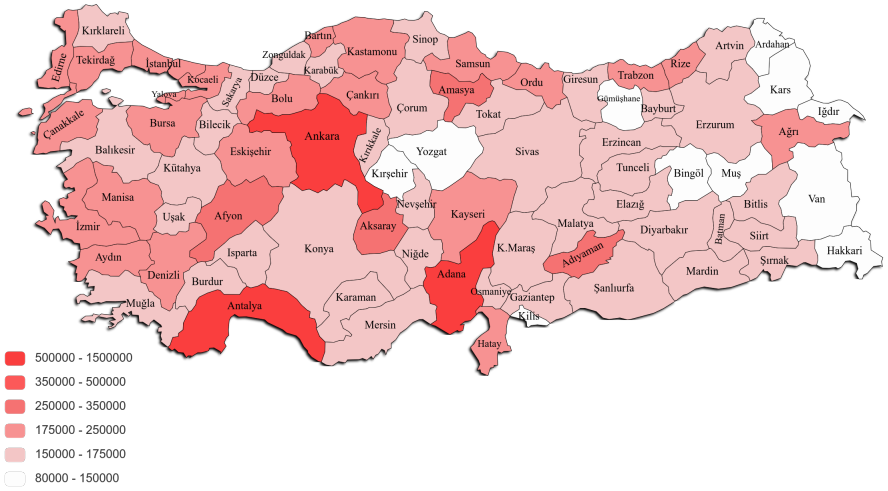
4.3.2 Veri ön işleme

Twint kütüphanesi ile tweetlerin tüm alanları çekilmiştir. Her tweetin 36 farklı bilgi alanı bulunmaktadır. Toplam veri bu alanlarla birlikte 7.6 GB boyutundadır. Bu çalışmada sadece metin verisi kullanıldığı için kalan 35 field silinmiştir.

Metin verisinde tab, new line gibi tüm boşluklar space karakteri ile değiştirilmiştir. Veriden mention, url ve hashtagler silinmiştir. Ardından her il için bir adet txt dosyası olmak üzere 81 adet txt dosyasına her bir satıra bir tweet gelecek şekilde metinler atılmıştır. Toplam veri boyutu 81 il için 1.5 GB olmuştur. Yapılan deneylerde farklı miktarda eğitim verisi kullanılmıştır. Tüm veri üzerinden filtrelenerek seçilmiştir. Deneyler kısmında veri filtreme yöntemi anlatılmaktadır.

4.3.3 Veri istatistiği

Her şehir için veri sayısının ısı haritası Şekil 4.2'de bulunmaktadır. En az tweet 88.390 adet Ardahan'da, en çok tweet 1.054.767 adet Ankara'da bulunmaktadır. Toplam tweet sayısı 16.116.035 olduğu için bir şehirdeki ortalama tweet sayısı 198.964 olmuştur. Öncelikle hazırlanan scriptlerde her şehirden 1000 kullanıcı ve her kullanıcının tüm tweetleri çekilmiştir. Ardından bu işlem çok uzun sürdüğünden dolayı ve eğitim verilerinin filtrelenerek sınırlandırılacağı için bu sayı önceki bölümlerde bahsedilen şekle düşürülmüştür. Bu sebeple, alfabetik olarak önde olduğundan dolayı A harfiyle başlayan şehirlerin veri sayıları diğerlerine oranla daha yüksektir.



Şekil 4.2: Konum tespiti ısı haritası.

5. ÖNERİLEN YÖNTEMLER

Bu bölümde incelenen tüm görevler için önerilen yöntemler detaylandırılacaktır.

5.1 Kontrole Değer İddiaların Tespiti

Bu bölümde kontrole değer iddiaların tespiti için kullanılan 5 farklı yöntemin detayları açıklanacaktır.

Dile Özel Eğitim: Önceki çalışmalar, tek bir dil için önceden eğitilmiş ince ayarlı dönüştürücü dil modellerinin çeşitli doğal dil işleme görevlerinde yüksek performans sağladığını ve en iyi sonuçları geride bıraktığını göstermiştir. Bu nedenle bu yöntemde her dil için dile özgü bir dönüştürücü model kullanılmıştır. Sırasıyla Türkçe için BERTurk[73], Arapça için AraBERT [4], İspanyolca için BETO [9], Bulgarca için RoBERTa [46] ve İngilizce için BERT [17] temel modeli kullanılmıştır. Her modele ilgili eğitim verileriyle ince ayar yapılmıştır. Bu yöntemde, iki farklı yaklaşımı incelenmiştir: 1) orijinal tweet'ler ($LSM_{orijinal\_tweetler}$) kullanılarak ince ayar yapılmış dile özgü modeller ve 2) temizlenmiş tweet'ler ($LSM_{temiz\_tweetler}$) kullanılarak ince ayar yapılmış dile özgü modeller. İkinci yöntemde tweet'lerden tüm bahsetmeler ve URL'ler kaldırılmıştır.

Veri Dağılımını Eşitleme: Rastgele bir tweet örneğinde, kontrol edilmeye değer iddialarla karşılaşma olasılığı düşüktür ve bu da dengesiz veri dağılımına neden olmaktadır. Bu durum görev için paylaşılan veri kümelerinde de gözlemlenmektedir. (Bkz. Çizelge 4.1). Eğitim setindeki kontrole değer tweetlerin oranı Arapça, Bulgarca, İngilizce, İspanyolca ve Türkçe için sırasıyla %22, %13, %35, %8 ve %38 şeklindedir.

Dengesiz etiket dağılımı, modeller için öğrenme sürecini olumsuz etkilemektedir. Veri kümesini tamamen dengeli hale getirmek için, kontrol edilmeye değer iddiaları aşırı örneklenebilir veya kontrole değer olmayan iddialar alt örneklenebilir. Yasser ve diğ. [88], aşırı örneklemenin, modellerinin kontrol edilmeye değer tespiti üzerindeki performansını iyileştirmediğini bildirmiştir. Bu nedenle, diller arasında adil bir karşılaştırma yapmak için tüm diller için aynı kontrole değer iddia oranını ayarlayarak alt örnekleme yaklaşımı sunulmuştur.

Ancak, her dilde aynı oranda etiket olması için birçok tweet'in kaldırılması gerekmektedir. Bu nedenle, alt örnekleme ile eğitim setinin kontrole değer talep oranı %30 yapılmıştır. Özellikle, Arapça, Bulgarca ve İspanyolca veri kümelerinde kontrol edilmeyen iddialar alt örneklenmiştir. Türkçe ve İngilizce veri setleri için kontrole değer iddia oranı %30'dan fazladır. Bu nedenle, bu veri kümelerinde kontrol edilmeye değer iddialar alt örneklenmiştir. Pozitif sınıfı alt örneklemek etkili olmayabilir, ancak diller arasında adil bir karşılaştırma yapmak için gerekmektedir.

Alt örnekleme yaparken veriler rastgele silinmiştir. Burada çeşitli yöntemler uygulanabilir. Örneğin, birbiriyle benzer metinler bulunup, benzerlik oranı en yüksek olanlar silinebilir. Ayrıca, modele etkisi en düşük olan veriler silinebilir. Alt örnekleme yapıldıktan sonra, bahsedildiği gibi her dil için bahsetmeleri ve URL'leri kaldırılmıştır ve dile özgü modellerde ince ayar yapılmıştır.

Makine Çevirisi: Etiketlenmiş veri miktarı, eğitilen modeller üzerinde önemli bir etkiye sahiptir. Ancak etiketleme, maliyetli ve zaman alıcı bir süreçtir. Bu nedenle, etiketlenmiş veri boyutunu otomatik olarak artırmak için makine çevirisi yöntemlerinden yararlanılmıştır. Özellikle her dil için diğer dillerde kontrol edilmeye değer olarak etiketlenen tweet'ler Google Translate kullanarak çevrilmiştir. Çeviriden sonra, bahsedenler ve URL'ler kaldırılmıştır ve her dil için dile özgü modeller ince ayar yapılmıştır. Bu yöntem aynı zamanda dengesiz etiket dağılımı sorununu da azaltmıştır. Örneğin, İspanyolca veri seti için kontrole değer iddia oranı bu yöntemle %50,8'e yükselmektedir.

Zayıf Denetim: Etiketlenmiş veri boyutunu artırmanın başka bir yolu da zayıf denetim [66]'dir. Bu nedenle, aşağıdaki zayıf denetim yöntemi kullanılmıştır. Önce her veri kümesindeki kelimeler sıklıklarına göre sıralanmıştır ve en sık kullanılan 100 kelime arasından ilgili veri kümelerinin konusuyla ilgili 10 kelime manuel olarak seçilmiştir. Bu kelimelerin her biriyle 500'er ve toplam her dil için 5000'er olmak üzere veriler çekilmiştir. Bu verinin detaylarına 4.1.2 bölümünden ulaşılabilir. Ardından, toplanan bu tweet'ler, ilgili eğitim verilerinin temizlenmiş tweet'leri kullanılarak ince ayar yapılan XLM-R [14] modeli kullanılarak etiketlenmiştir. Son olarak, tweet'lerden URL'ler ve bahsetmeler kaldırılmıştır ve asıl eğitim verileriyle birleştirilerek her dil için dile özgü dönüştürücü modeller ile ince ayar yapılmıştır.

Çapraz Dilli Eğitim: Çok dilli dönüştürücü modeller farklı dillerin verileriyle birlikte eğitildiği için, etiketli veri bulunan bir dil ile ince ayar yapılarak, farklı dillerde sonuç almaya olanak sağlamaktadır. Bu nedenle, özellikle düşük kaynaklı diller için doğal dil işlemede büyük potansiyele sahiptirler. Bu yöntemde, diller arası eğitimin kontrol edilmeye değer iddiaları tespit etmede etkili olup olmadığı araştırılmıştır. Özellikle, her dil çifti için eğitim verileri birleştirilir ve birleştirilmiş veri seti kullanılarak mBERT [17] modeli ince ayar yapılmıştır. Daha sonra, ince ayarlı model beş dilin tamamında test edilmiştir.

5.2 Taraf Tespiti

Önceden eğitilmiş dil modelleri, son derece büyük veri kümeleri ve çok sayıda parametre kullanarak dillerin genel özelliklerini öğrenmektedir. Ardından, modellerin göreve özel ayrıntıları öğrenmesi için etiketlenmiş verilere ihtiyaç bulunmaktadır. Bu nedenle, veri boyutu ve etiket kalitesi gibi etiketlenmiş veri kümelerinin özellikleri, onlarla eğitilen modeller üzerinde büyük etkiye sahiptir. Bu çalışmada, taraf tespiti için eğitim verilerinin etkisi domain bilgisi, tarafın hedefi, kullanılan dil ve veri boyutu olmak üzere dört açıdan incelenmiştir.

Domain Bilgisi: Belirli bir hedefe yönelik tarafı ifade eden ifadeler ve kelimeler, hedefin domainine göre değişmektedir. Örneğin görev, insanların (örneğin politikacılar) insanlara karşı tarafını tespit etmekse, "taraftar" ve "düşman" gibi kelimeler kullanı-

labilir. Ancak hedef bir fikirse, örneğin iklim değişikliği ise, bu kelimeleri tarafı ifade etmek için kullanmak pek mümkün değildir. Benzer şekilde Türkçe'de de farklı alanlarda kullanılan ortak ifadeler farklılık göstermektedir. Örneğin "tutmak" normal anlamıyla "tutmak", mecazi anlamıyla ise "desteklemek" anlamına gelir. Spor kulüplerine halkın desteğini ifade eden "taraftar olmak" ile "tutmak" ortak ifadelerdir. Ancak siyasi alanda bu sözler çok güçlü bir tarafa işaret etmektedir. Bu nedenle "seçmen" ve "desteklemek" gibi ifadeler daha çok kullanılmaktadır. Çalışmada, domain bazlı taraf tespiti için hassas ayarlanmış BERT modellerinin performansı üzerindeki etkisini araştırmak için spor, sağlık, politika ve ekonomi gibi farklı alanlardan gelen veriler kullanılmıştır.

Tarafın Hedefi: Daha önce bahsedildiği gibi, mevcut birçok veri kümesinin önceden tanımlanmış bir hedefi vardır ve görev, hedefin tespit edilmesini içermez. Ancak bu, farklı hedeflerde taraf tespiti için mevcut veri kümelerini kullanılamaz hale getirmektedir. Bunun nedeni, bir tarafı ifade eden bazı ifadelerin hedefe özel olmasıdır. Örneğin, Liverpool Futbol Kulübü taraftarları kendilerine "kopites", Machester United taraftarları ise genellikle "citizens" olarak adlandırmaktadır. Beşiktaş Futbol Kulübü taraftarları için Türkçe'de "Çarşı" gibi benzer ifadeler bulunmaktadır. Bu nedenle, hassas ayar yapılan bir model, hedefe özel ifadeleri öğrenip genelleştirebilir. Çalışmada, modelleri eğitmek için tek hedefli, çok hedefli ve çapraz hedefli veri kullanmanın etkisi incelenmiştir.

Kullanılan Dil: Doğal dil işlemede birçok çözüm dile özgüdür. Ancak MBERT ve XLMR gibi çok dilli dönüştürücü dil modelleri, eğitim ve test veri kümelerinin farklı dillerde olabileceği çözümler geliştirmek için heyecan verici bir alternatif yaklaşım sunmaktadır. Bu özellik, İngilizce olmayan diller için doğal dil işleme görevlerinde daha fazla ilerleme sağlamak için büyük bir potansiyele sahiptir. Bunun nedeni, dile özgü bir model kullanmanın, bir dilin dilsel ayrıntılarını daha iyi öğrenmesi nedeniyle çok dilli modellerden daha iyi performans göstermesidir. Sınırlı veri kümelerine sahip olmak, düşük veya orta kaynaklı diller için doğal dil işleme modelleri geliştirmenin önündeki ana engellerden biridir. Örneğin taraf tespiti için iki futbol kulübünün tweetlerini kapsayan tek bir [41] veri seti bulunmaktadır. Bu nedenle, belirli bir dilin sınırlı kaynaklarını kullanmak yerine, başka bir dilin zengin kaynaklarını kullanmak daha faydalı olabilir. Çalışmada Türkçe taraf tespiti görevi için İngilizce veri setleri kullanmanın sonucu incelenmiştir.

Veri Boyutu: Etkili eğitim için büyük miktarda etiketlenmiş veriye ihtiyaç olsa da, manuel etiketleme yavaş ve maliyetlidir. Bu nedenle, etiketleme bütçesini optimize etmek için kitle kaynaklı etiketleme [70] ve aktif öğrenme [67] gibi çeşitli yöntemler incelenmiştir. Bununla birlikte, bu yaklaşımlar hala önemli miktarda para ve zaman gerektirir. Etiketlenmiş veri boyutunu artırmaya yönelik bir başka popüler yaklaşım, verilerin başka bir model kullanılarak otomatik olarak etiketlendiği zayıf denetimdir. Zayıf denetim yöntemlerinde, gürültülü etiketlere sahip büyük verilerin, doğru yargılara sahip sınırlı verilerden daha faydalı olması beklenmektedir. Çalışmada, otomatik olarak etiketlenen verilerin hassas ayar taraf tespiti modelleri üzerindeki etkisini incelenmiştir.

5.3 Konum Tespiti

Bu bölümde konum tespiti modellerinin başarımını etkileyebilecek veri kümesi özellikleri ile eğitim verisi oluşturma yöntemleri anlatılacaktır. Diğer görevlerde yalnızca dönüştürücü dil modeli kullanılırken, bu görevdeki tüm yöntemler MLP, GCN ve BERT olmak üzere üç farklı modelle incelenmiştir.

Veri Kümesi Büyüklüğü: Coğrafi konum tespiti çalışmalarında kullanıcı profilinin meta bilgileri ve metin paylaşımları kullanılmaktadır. Fakat kullanıcı bilgilerini toplamak daha zor olduğu için çalışmalarda daha çok metin verisi kullanılmıştır. Metin verisi daha çok olduğu için konum tespit çalışmalarında önemli bir rol oynamaktadır. Bu yöntemin amacı kullanıcının paylaşım miktarının konum tespitindeki etkisini öğrenmektir.

Bu yöntemde her kullanıcının paylaşım sayısının konum tespitine oranı incelenmiştir. Temel karşılaştırmada her kullanıcıdan 100 tweet alınırken, bu yöntemde tüm algoritmalar için 75 tweet, 50 tweet ve 25 tweet alınarak sonuçlar incelenmiştir.

Konuya Özel Veri Filtreleme: Kullanıcılar sosyal medyada her alandan çeşitli paylaşımlar yapmaktadır. Bu paylaşımların çoğu konum ile bağlantısız veriler olabilir. Bu sebeple konum tespiti yaparken metin verisini filtrelemek ve sadece konum ile ilgili olanları kullanmak iyi bir yöntem olabilir.

Veri kalitesinin konum tespitine etkisini ölçmek için, tüm kullanıcıların tüm tweetlerini kullanmak yerine yalnızca belirlenen koşulları sağlayan metinler filtrelenmiştir. Veri setinin alt kümeleri kullanılarak, tüm parametre ve modeller sabit tutularak yeniden deney yapılmıştır. Kullanıcıların son 100 tweeti yerine, önceden seçilen kelimeleri içeren tweetlerin en fazla 100 tanesi seçilerek yapılmıştır. Tweetler filtrelendiğinde normalden daha az veri gelmektedir. Fakat alınan tüm verilerde spesifik kelimeler bulunmaktadır.

Kelime listesine öncelikle 81 il ismi ve tüm ilçe isimleri eklenmiştir. Ardından Türkçe kelime gömme modeli kullanılarak, her il ismine en benzer 3 kelime alınmıştır. Bu 3 kelime de listeye eklenmiştir. Ardından aynı kelimeler silinmiş ve liste oluşturulmuştur.

6. DENEYLER

Bu bölümde her bir konu için ayrı ayrı deney düzeneği, implementasyon ve deney sonuçları alt başlıklar halinde açıklanacaktır.

6.1 Kontrole Değer İddiaların Tespiti Deneyleri

Kontrole değer iddiaların tespiti çalışması, Check That21 labında yapılmıştır. Öncelikle yarışma sürecinde verilen geliştirme seti kullanılarak deneyler yapılmıştır ve test set sonuçları gönderilmiştir. Bu çalışmada test set üzerindeki sonuçlara değinilecektir. Tüm deney detaylarına Zengin ve diğ. [91] çalışmasından ulaşılmaktadır.

6.1.1 Deney düzeneği

Modellere ince ayar yapmak için ktrain [48] kütüphanesi kullanılmıştır. Her modelin parametrelerini ayarlamak için çeşitli konfigürasyonlara sahip geliştirme seti üzerinde deneyler yapılmıştır ve her model için en iyi performans gösteren seçilmiştir. Kullanılan her modelin parametreleri Çizelge 6.1'de bulunmaktadır. Tüm modellerin öğrenme oranı 5e-5 olarak belirlenmiştir. Tüm sonuçlarda ortalama hassasiyet puanı kullanılmıştır.

Çizelge 6.1: Kullanılan önceden eğitilmiş model isimleri.

Dil	**Model İsmi**	**Çalıştırma Boyutu**	**Epoch**
Arapça	AraBERT	6	1
Bulgarca	RoBERTa Base for Bulgarian	6	3
İngilizce	BERT Base	3	3
Çok Dilli BERT	mBERT	6	1
İspanyolca	BETO	6	1
Türkçe	BERTurk	6	3

6.1.2 Deney sonuçları

Kontrole değer iddiaların tespiti çalışmasının tüm deney sonuçları Çizelge 6.2'de bulunmaktadır. $LSM_{orijinal\_tweet}$ dile özel önceden eğitilmiş model ve orijinal metin içeriği bulunmaktadır. LSM_{temiz_tweet} dile özel önceden eğitilmiş model ve temizlenmiş metin içeriği sonuçları bulunmaktadır. Çapraz dilli eğitim sonuçlarında ise tüm diller için bulunan en yüksek sonuçlar yazılmıştır. Bulgarca için TR+BG tweetleri, İngilizce ve Türkçe için TR+İNG tweetleri, Arapça için TR+AR tweetleri ve İspanyolca için TR+İSP tweetleri ile eğitilmiş mBERT modelinin sonuçları bulunmaktadır.

Çizelge 6.2: Test setinde ortalama hassasiyet puanı.

Yöntem İsmi	**Arapça**	**Bulgarca**	**İngilizce**	**İspanyolca**	**Türkçe**
$LSM_{orijinal\_tweet}$	0.600	**0.548**	**0.172**	0.505	0.553
LSM_{temiz_tweet}	0.575	0.149	0.081	**0.537**	**0.581**
Alt Örnekleme	**0.622**	0.241	0.158	0.522	0.580
Zayıf Denetim	0.546	0.217	0.156	0.489	0.535
Makine Çevirisi	0.524	0.228	0.126	0.457	0.489
Çapraz Dilli Eğitim	0.543	0.532	0.151	0.149	0.443
CTL21 En İyi Sonuç	0.658	0.737	0.224	0.537	0.581
CTL21 İkinci Sonuç	0.615	0.673	0.195	0.529	0.574

Bu sonuçlara göre, Arapça dilinde aşırı örnekleme yaparak veri dağılımını düzenleme yöntemi başarılı sonuç vermiştir. Bulgarca ve İngilizce dillerinde orijinal metinleri kullanarak dile özel dönüştürücü dil modeline hassas ayar yapmak en iyi sonucu vermiştir. İspanyolca ve Türkçe dillerinde ise metinleri ön işlemden geçirmek ve dile özel dönüştürücü dil modeline hassas ayar yapmak en iyi sonucu vermektedir.

İngilizce dilinde ortalama hassasiyet puanı diğer dillere göre oldukça düşük çıkmıştır. Bunun sebebi eğitim ve test setindeki etiket dağılımının oldukça farklı olması olabilir. En alt iki satırda bulunan CTL21'deki en iyi sonuçlarla kıyaslandığında katılımcıların diğer dillere göre düşük skor aldığı görülmektedir. Tüm katılımcılarda düşük skor çıkması etiket dağılımının dengesizliğinden olduğunu göstermektedir. Bununla birlikte Bulgarca dilinde, orijinal tweet ile dile özel eğitim yapıldığında 0.548 sonuç alınırken, temizlenmiş tweet kullanıldığında 0.149 sonuç alınmıştır. Veri incelendiğinde buradaki farkın sebebinin temizleme yapılırken silinen hashtag'lerin veri dağılımındaki oranının farklı olduğu olarak düşünülmüştür. Kontrole değer iddiaların içerisinde %43.6 hashtag oranı varken, kontrole değer olmayan metinlerdeki oran %82.6'dır.

Çizelde 6.3'de geliştirme setinde yapılan çapraz dilli eğitim kombinasyonlarının sonuçları yer almaktadır. Bu kombinasyonlar birli, ikili, üçlü, dörtlü, beşli olacak şekilde tüm ihtimallerle denenmiştir. Çizelgede yalnızca ikili sonuçlar yer almaktadır. Bu sonuçlara göre, Arapça dışındaki diğer dört dilde en iyi sonucu almak için Türkçe dilini eğitim verisi olarak kullanmak gerektiği çıkarılmaktadır. Arapça dilinde ise İngilizce kullanmak başarılı sonuç vermiştir. Ayrıca tüm dillerde en iyi sonucu almak için, beklendiği üzere eğitim verisi olarak kendi dilinin kullanılması gerekmektedir.

Ayrıca beş farklı dil için test setinde alınan en yüksek sonuçlar CTL21'de paylaşılmıştır. Resmi sonuçlara deneylerde alınan en iyi sonuçlar Türkçe ve İspanyolca dillerinde 1. sırada almaktadır. Model sonuçları Bulgarca'da 4, Arapça'da 6 ve İngilizce'de 10. sırada yer almıştır.

Çizelge 6.3: Geliştirme setinde çapraz dilli eğitim puanı.

Eğitim Dili	Arapça	Bulgarca	İngilizce	İspanyolca	Türkçe
AR + İSP	0.349	0.274	0.457	0.118	0.451
BG + AR	0.512	0.194	0.481	0.077	0.372
BG + İSP	0.386	0.362	0.477	0.187	0.522
İNG + BG	0.252	0.152	0.536	0.140	0.511
İNG + AR	**0.713**	0.241	0.564	0.156	0.505
İNG + İSP	0.183	0.195	0.410	0.269	0.417
TR + BG	0.433	**0.505**	0.607	0.121	0.585
TR + İNG	0.532	0.264	**0.610**	0.135	**0.601**
TR + AR	0.606	0.214	0.507	0.090	0.536
TR + İSP	0.300	0.189	0.495	**0.298**	0.556

6.2 Taraf Tespiti Deneyleri

Bu bölümde taraf tespiti çalışmasında yapılan deneyler anlatılmaktadır. Sırasıyla deney düzeneği, deney sonuçları ve model açıklanabilirliği anlatılmaktadır.

6.2.1 Deney düzeneği

Huggingface [85]'den alınan önceden eğitilmiş dil modelleri kullanılmıştır. Modellere hassas ayar yapmak için Ktrain kütüphanesi [48] kullanılmıştır. Öğrenme oranı, maksimum uzunluk, parti boyutu ve epoch sayısı sırasıyla 5e-5, 500, 5 ve 3 olarak belirlenmiştir. Rastgele tohum değerleri olarak üç farklı değer kullanılmıştır ve değerler arasında önemli bir değişiklik gözlemlenmemiştir. Bu nedenle, adil bir karşılaştırma yapmak için tüm deneyler için 42 seçilmiştir. Türkçe veriler için BERTurk [73] ve İngilizce veriler ve diller arası deneyler için M-BERT [17] kullanılmıştır. Eğitim ve test seti oranı %90 ve %10 olarak ayarlanmıştır. Değerlendirme için makro ortalama F1 puanı raporlanmaktadır.

6.2.2 Deney sonuçları

İlk deney, aynı domainde olan dört farklı futbol takımı kullanılarak yapılmıştır. Burada eğitim verisinde bulunan hedef değiştirilerek, farklı hedeflere ait sonuçlar alınmıştır. Tek hedefli ve çok hedefli sonuçlar alınmıştır. Çapraz hedef tespiti sonuçları Çizelge 6.4'de bulunmaktadır.

Bu sonuçlara göre, tek hedefli taraf tespiti deneylerinde Trapzonspor takımı hariç en yüksek skor, tüm veri kullanılarak hassas ayar yapılan modelde alınmıştır. Trabzonspor için alınan en yüksek skor ise kendi verisi ve Beşiktaş'ın birleştiği model olmuştur. Çok hedefli testte de en yüksek skor, tüm eğitim verisi birleştirildiği zaman alınmıştır.

Yalnızca tek bir takımın eğitim verisi birleştirildiğinde ve tek hedefli test yapıldığında, ortalama en yüksek skor Galatarasay ile eğitilen modelde alınmıştır. Çünkü veri sayısı diğer takımlara göre daha yüksektir. Yani bu deneyden çıkarılan sonuç, kaliteli, büyük

Çizelge 6.4: Çapraz hedef sonuçları.

	Tek Hedefli Test				**Çok Hedefli Test**
Eğitim Verisi	GS	FB	BJK	TS	GS+FB+BJK+TS
GS	0.633	0.695	0.650	0.587	0.640
FB	0.631	0.589	0.413	0.492	0.585
BJK	0.245	0.294	0.576	0.492	0.384
TS	0.419	0.711	0.529	0.689	0.567
GS + FB	0.834	0.639	0.700	0.487	0.744
BJK + TS	0.705	0.677	0.740	**0.849**	0.708
GS+FB+BJK+TS	**0.890**	**0.750**	**0.879**	0.707	**0.823**

miktarda ve tek hedefli bir veri bulunuyorsa, bu veri aynı domainde farklı hedefler için kullanılabilir. Ayrıca farklı hedef verileri varsa bunları birleştirmek ve çok hedefli bir model eğitmek başarıyı arttırmaktadır.

Eğitim verisini zayıf denetim ile arttırdıktan sonra alınan sonuçlar Çizelge 6.5'de bulunmaktadır. Zayıf denetim olarak otomatik etiketleme yapılmıştır. Twitter üzerinden belirlenen kelimelerle çekilen metinler etiketli veri olarak varsayılmış ve bu verilerle hassas ayar yapılmıştır. Eğitim verisi olarak öncelikle takımların teker teker otomatik etiketli metinleri kullanılmıştır. GS_o, FB_o, BJK_o, TS_o her takımın otomatik etiketli veri kümesini temsil etmektedir. $Hepsi_o$ ise tüm takımların tüm otomatik etiketli verisini temsil etmektedir. GS_m, FB_m, BJK_m, TS_m, her takımın manuel olarak etiketlenmiş verilerini temsil etmektedir. $Hepsi_m$ ise tüm takımların manuel etiketli veri kümsesini temsil etmektedir.

Çizelge 6.5: Zayıf denetim sonuçları.

Eğitim Verisi	**GS**	**FB**	**BJK**	**TS**	**GS+FB+BJK+TS**
GS_o	0.444	0.714	0.607	0.750	0.596
FB_o	0.474	0.714	0.560	0.654	0.569
BJK_o	0.289	0.369	0.354	0.514	0.411
TS_o	0.303	0.552	0.457	0.786	0.544
$Hepsi_o$	0.303	0.294	0.422	0.823	0.535
$GS_m + GS_o$	0.638	0.428	0.777	0.776	0.696
$FB_m + FB_o$	0.515	0.779	0.601	0.805	0.652
$BJK_m + BJK_o$	0.430	0.374	0.419	0.196	0.301
$TS_m + TS_o$	0.572	0.576	0.798	0.823	0.668
$Hepsi_m + Hepsi_o$	**0.860**	**0.818**	**0.870**	**0.833**	**0.870**

Bu sonuçlara göre, az veri kullanılarak yapılan taraf tespit çalışmalarında otomatik etiketleme ile veri arttırmanın model başarısını arttırdığı görülmektedir. Yalnızca otomatik etiketleme ile eğitilen modeller yeterli performansı göstermemektedir. Fakat veriyi manuel etiket ile birleştirmek en iyi sonucu vermiştir. Ayrıca takım bazında incelendiğinde ise, bir önceki deneyde olduğu gibi verisi çok olan takım en yüksek başarıyı vermektedir. Manuel etiketlenen tüm veri ve otomatik etiketlenen tüm veri birleştirildiğinde en yüksek başarı çıkmaktadır.

Farklı domainler arasındaki yapılan test sonuçları Çizelge 6.6'de bulunmaktadır. Futbol, sağlık, ekonomi ve siyaset domainleri kullanılmıştır. Her domain teker teker hassas ayar yapılmış ve diğer tüm domainlerde test edilmiştir.

Çizelge 6.6: Çapraz domain deney sonuçları.

Domain	Futbol	Sağlık	Ekonomi	Siyaset
Futbol	**0.793**	0.610	0.556	0.476
Sağlık	0.548	**0.864**	0.700	0.665
Ekonomi	0.597	0.666	**0.912**	0.590
Siyaset	0.511	0.504	0.597	**0.771**

Çapraz domain deney sonuçlarına göre, beklendiği gibi her domainde alınan en yüksek skor, kendi domain verisi kullanılarak eğitilen modelde alınmıştır. Yani taraf tespit çalışmalarında belli bir domaine özel hazırlanacak modelin, kendi domain verisi kullanılan verilerle eğitilmesi gerekmektedir. Farklı domainlerde bulunan verilerle yapıldığı zaman yeterli performans alınamamaktadır.

Çapraz dil deney sonuçlarına Çizelge 6.7'de bulunmaktadır. Burada model verisi olarak öncelikle İngilizce veri kümeleri kullanılmıştır. Ardından İngilizce veri kümeleri ve aynı domaindeki Türkçe veri kümeleri birleştirilerek modeller eğitilmiştir. Tüm test kümeleri Türkçe'de farklı domaindeki veriler olmuştur. $politika_e$, $feminizm_e$, $iklim_e$, $kurtaj_e$, $ateizm_e$, $saglik_e$, $ekonomi_e$, $siyaset_e$ İngilizce veri kümelerini ifade etmektedir. $saglik_t$, $ekonomi_t$, $siyaset_t$ ise Türkçe veri kümelerini ifade etmektedir.

Çizelge 6.7: Çapraz dil deney sonuçları.

Eğitim Verisi	**Futbol**	**Sağlık**	**Ekonomi**	**Siyaset**
$siyaset_e$	0.566	0.629	0.518	0.466
$feminizm_e$	0.381	0.486	0.582	0.296
$iklim_e$	0.288	0.201	0.133	0.372
$kurtaj_e$	0.408	0.454	0.582	0.296
$ateizm_e$	0.483	0.515	0.569	0.380
$saglik_e$ + $saglik_t$	0.503	**0.765**	0.657	0.358
$ekonomi_e$ + $ekonomi_t$	0.556	0.381	**0.588**	0.296
$siyaset_e$ + $siyaset_t$	**0.680**	0.605	0.583	**0.756**

Diller arası test sonuçlarına göre, farklı dillerde düşük miktarda eğitim verisi kullanarak hassas ayar yapıldığında yeteri kadar performans alınmamaktadır. Fakat düşük miktardaki farklı dil eğitim verisi ile aynı dildeki eğitim verisi birleştirildiğinde performans artmaktadır. Burada kullanılan model, çok dilli dönüştürücü dil modeli olduğu için, iki farklı dildeki eğitim verisini birleştirmek Türkçe dilinde daha başarılı sonuçlar vermiştir.

Sonuç olarak, bir hedef özgü taraf tespiti yapılıyorsa, belirli bir hedef için verilerin etiketlenmesine ve o hedef üzerinde sonuç alınmasına gerek yoktur. Hedeften bağımsız etiketleme yapmak veya hedefleri birleştirerek daha büyük bir veri ile hassas ayar yapmak daha başarılı sonuçlar vermektedir. Bu sonuç futbol veri setinin deney sonuçlarından çıkarılmaktadır. Örneğin Galatasaray ve Fenerbahçe için veri etiketi yapılırsa bu düşük boyutlu veriler ile taraf tespitinin herhangi bir takım için birleştirilmesi

mümkündür. Ancak sadece bir takım kullanılıyorsa ve veri boyutu küçükse hem kendi takımında hem de diğer takımlarda daha düşük sonuçlar vermektedir.

Yalnızca otomatik etiketleme yaparak model eğitmek başarılı sonuçlar vermemektedir. Fakat otomatik etiketleme ile hazırlanan veriler, az sayıda etiketlenmiş veri ile birleştiğinde daha başarılı sonuçlar alınmaktadır.

Taraf tespiti farklı domainlerde yapıldığında, her domainin kendine özel veri setinin bulunması önerilmektedir. Her domaine özel model eğitmek gerekmektedir. Çünkü futbol için eğitilen taraf tespit modeli, siyaset domaininde başarılı sonuç vermemektedir. Çünkü siyaset için destekleyici ve karşıt ifadeler, futbol domainindekilerden farklıdır. Çapraz domain sonuçlarına göre her domainde sadece kendi domain verileri ile eğitildiğinde en iyi sonuçlar alınmaktadır. Diğer bir deyişle, tüm alanlarda ve tüm hedeflerde başarılı bir şekilde çalışan taraf tespit modeli mümkün değildir. Ancak domaine özel taraf tespiti başarılı bir şekilde yapılmaktadır.

Son olarak, çapraz dil deney sonuçlarına göre, İngilizce veriler kullanılarak çok dilli olarak eğitilen taraf tespit modelleri Türkçe'de yeterince başarılı değildir. Manuel etiketlenen az sayıdaki Türkçe veri ile İngilizce veri birleştirildiğinde, sonuçlar sadece Türkçe veri kullanılarak alınana göre daha düşüktür. Yani verileri farklı bir dille artırmak, Türkçe için performansı arttırmamıştır.

6.2.3 Model açıklamaları

Önceden eğitilmiş dil modelleri, sınıflandırma görevlerinde oldukça iyi performans gösterir. Ancak bir dezavantaj olarak şeffaflıktan yoksundur. Bu nedenle, tahminleri açıklamaya yardımcı olan açık kaynaklı bir Eli5 kütüphanesi kullanarak LIME [68] algoritması ile modeller incelenmiştir. Buradaki sonuçlar, modeller tahminlerinin doğru veya yanlış olmasına sebep olan kelimeleri ve nedenleri göstermektedir. Çapraz dil deneyleri ve çapraz domain deneyleri için model açıklamaları incelenmiştir.

Şekil 6.1'de çapraz domain açıklamaları bulunmaktadır. Model olarak siyaset, ekonomi, futbol ve sağlık domainleri kullanılmıştır. Cümle olarak "mali açıdan zor durumda olan takımımızın çok fazla sakat oyuncusunun olması başkanlık seçimi öncesinde takımı zor duruma sokacağı için başkanımızın yanındayız" seçilmiştir. Bu cümle domain olarak siyaset ve futbol içermektedir. Cümlede hem futbol domaini için hem de siyaset domaini için destekleyici bir tavır bulunmaktadır.

Model	Kelime Önemi	Tahmin	Gerçek Değer
Siyaset	mali açıdan zor durumda olan takımımızın çok fazla sakat oyuncusunun olması başkanlık seçimi öncesinde takımı zor duruma sokacağı için başkanımızın yanındayız.	**Destekleyici** Olasılık **0.968**	Destekleyici
Ekonomi	mali açıdan zor durumda olan takımımızın çok fazla sakat oyuncusunun olması başkanlık seçimi öncesinde takımı zor duruma sokacağı için başkanımızın yanındayız.	**Karşıt** Olasılık **0.611**	
Futbol	mali açıdan zor durumda olan takımımızın çok fazla sakat oyuncusunun olması başkanlık seçimi öncesinde takımı zor duruma sokacağı için başkanımızın yanındayız.	**Destekleyici** Olasılık **0.751**	
Sağlık	mali açıdan zor durumda olan takımımızın çok fazla sakat oyuncusunun olması başkanlık seçimi öncesinde takımı zor duruma sokacağı için başkanımızın yanındayız.	**Karşıt** Olasılık **0.983**	

Şekil 6.1: Çapraz domain açıklamaları.

Sonuçlar incelendiğinde siyaset ve futbol modelleri, siyaset ve futbol konularının geçtiği test cümlesi için doğru sonuçlar vermektedir. İkisinde de destekleyici olduğunu tahmin etmiştir. Ancak alakasız olan ekonomi ve sağlık modelleri yanlış sonuçlar vermiştir. Tahminlerdeki en önemli kelimenin “yanındayız" olduğu görülmektedir. Yani farklı domainlerde "yanındayız" kelimesi karşıt bir tavıra yol açmaktadır, fakat aynı domain olduğu zaman doğruya sevk etmektedir.

Şekil 6.2’de çapraz dil açıklamaları bulunmaktadır. MBERT modeli öncelikle Türkçe ekonomi veri seti ile, ardından İngilizce ekonomi veri seti ile hassas ayar yapılmıştır. Bu iki modele de hem Türkçe, hem İngilizce test cümlesi sorulmuştur. Test için ekonomi ve bitcoin konusunu içeren "projeye bakabilirsiniz, yatırım tavsiyesi değil ama bana göre güzel bir proje #bitcoin" metni seçilmiştir. İngilizce testte ise aynı cümlenin İngilizce çevirisi kullanılmıştır. Cümlenin gerçek etiketi destekleyici olarak belirlenmiştir.

Veri Seti	Kelime Önemi	Tahmin	Gerçek Değer
Türkçe Ekonomi	projeye bakabilirsiniz, yatırım tavsiyesi değil ama bana göre güzel bir proje #bitcoin	**Destekleyici** Olasılık **0.700**	Destekleyici
Türkçe Ekonomi	you can check the project, it's not investment advice, but it's a good project for me #bitcoin	**Destekleyici** Olasılık **0.836**	
İngilizce Ekonomi	projeye bakabilirsiniz, yatırım tavsiyesi değil ama bana göre güzel bir proje #bitcoin	**Karşıt** Olasılık **0.752**	
İngilizce Ekonomi	you can check the project, it's not investment advice, but it's a good project for me #bitcoin	**Destekleyici** Olasılık **0.566**	

Şekil 6.2: Çapraz dil açıklamaları.

Çapraz dil sonuçlarına göre, Türkçe veri seti ile MBERT modeline hassas ayar yapılan model, hem Türkçe hem de İngilizce testte doğru tahmin yapmıştır. Tahminin doğru olmasında "güzel bir proje", "a good project" gibi olumlu ifadeler etkili olmuştur. "değil" ve "not" gibi olumsuzlukların ise beklendiği üzere tam tersi etki yaptığı görülmektedir. İngilizce veri seti ile hassas ayar yapılan modelde ise, İngilizce test metni doğru, fakat Türkçe test metni yanlış tahmin edilmiştir. Burada "good project" ifadesi olumlu etki yaparken, "güzel bir" ifadesinin karşıt olmasında etkili olduğu görülmektedir.

6.3 Konum Tespiti Deneyleri

Bu bölümde konum tespiti için yapılan deneyler açıklanmaktadır. Öncelikle deney düzeneği ve kullanılan metrikler açıklanmıştır. Ardından deney sonuçları gelmektedir.

6.3.1 Deney düzeneği

Konum verileri Türkiye'nin 81 ili için şehir bazlı olarak çekilmiştir. Koordinat gereken algoritmalarda, şehir merkezlerinin koordinatı kullanılmıştır. Literatürdeki konum tespit çalışmalarında olduğu gibi, sistemlerin hatasını ölçmek için kullanıcıların gerçek konumu ile tahmin edilen konumu arasındaki mesafe kullanılmıştır. Daha özel olarak, ortalama uzaklık (mean), medyan uzaklık (median) ve doğruluk@161 (acc@161) metrikleri kullanılmıştır. Koordinatları verilen iki konum arasındaki mesafeyi ölçmek için aşağıda verilen haversine mesafe formülü kullanılmıştır.

$$\text{mesafe}(enl_1, boy_1, enl_2, boy_2) =$$
$$2r \arcsin \sqrt{\sin^2\left(\frac{enl_1 - enl_2}{2}\right) + \cos(enl_1)\cos enl_2 \sin^2\left(\frac{boy_1 - boy_2}{2}\right)}$$

Literatürde çalışmalar tüm kullanıcıların tam konumu üzerinde yapılmaktadır. Bu çalışmada kullanıcı bilgileri şehir bazlı tutulduğu için şehirlerin merkez koordinatları üzerinden hesaplanmıştır. GCN ve MLP modelleri literatürde olduğu gibi koordinat tabanlı çalışmaktadır. BERT modeli ise 81 il için sınıflandırma olarak çalıştırılmıştır. Burada sınıflandırma sonucunun başarısı F1 skoru ile değil, şehirlerin orta noktalarının koordinat uzaklığı kullanılarak yine aynı metriklerle ölçülmüştür.

Ortalama değer, tespit edilen konumların gerçek konumlara kilometre cinsinden ortalama uzaklığıdır. Medyan ise, tespit edilen konumların gerçek konumlara uzaklıklarının ortanca değeridir. Doğruluk@161 metriği ise tahmin edilen konum gerçek konumdan 161 km'den (yani 100 mil) daha yakın ise, tahminin doğru olduğu kabul edilerek hesaplanır. Doğruluk@161 aşağıdaki formül kullanılarak hesaplanmıştır.

$$\text{Doğruluk@161} = \sum_{i=0}^{n} \begin{cases} \mathbf{1} \text{ eğer } (enlem_i, boylam_i, enlem'_i, boylam'_i) < 161 \text{ km} \\ \mathbf{0} \text{ değilse} \end{cases}$$

n = kullanıcı sayısı
kullanıcının gerçek konumu = $(\text{enlem}_i, boylam_i)$
kullanıcının tahmin edilen konumu = $(enlem'_i, boylam'_i)$

Çalışmanın amacı dönüştürücü dil modellerine hassas ayar yapmak olduğu halde bu göreve özel MLP ve GCN modelleri de kullanılmıştır. Bunun sebebi birçok görevde dönüştürücü dil modelleri en yüksek sonuçları verirken, coğrafi konum tespiti görevinde MLP ve GCN modelleri yüksek performans göstermektedir. Bu sebeple BERT modeliyle birlikte toplam üç farklı model deneyi yapılmıştır.

Veri seti, tüm algoritmalarda ve yöntemlerde %60 eğitim, %20 geliştirme ve %20 test verisi olarak ayrılmıştır. Her ilden 400 kullanıcı olmak üzere, 81 ilden 32400 kullanıcı seçilmiştir. Ayrıca her kullanıcının tweet sayısı en fazla 100 olarak alınmıştır. BERT modeli için Türkçe BERTurk kullanılmış ve tüm yöntemlerde karşılaştırabilmek amacıyla aynı parametreler kullanılmıştır. 3 epoch, öğrenme oranı 5e-5, batch oranı 6, ve uzunluk olarak 500 kullanılmıştır.

Temel karşılaştırma için GCN, MLP ve BERT modelleri kullanılmıştır. GCN ve MLP algoritmaları Rahimi ve diğ. [65] çalışmasında açık kaynaklı olarak paylaşılan yöntem ile yapılmıştır. Temel karşılaştırma amacıyla tüm algoritmaların kullanılmasının sebebi, daha önce Türkiye'deki bir koordinat verisiyle konum tespiti yapılmamış olmasıdır. Burada alınan sonuçlar, sonraki yöntemlerde açıklanan sonuçlarla karşılaştırılmıştır.

6.3.2 Deney sonuçları

Daha önce belirtildiği üzere, konum tespiti amacıyla ne Türkiye coğrafi konumunda ne de Türkçe dilinde veri seti bulunmamaktadır. İngilizce dili için farklı coğrafi konumlara sahip metin ve ağ bilgisi bulunan GeoText[20], Twitter-US[69] ve Twitter-World[27] veri setleri bulunmaktadır. Türkiye konumunda ve Türkçe dilinde ilk defa sonuç alınmasına rağmen, karşılaştırma yapma amacıyla bu 3 İngilizce veri setinde metin tabanlı çalışma sonuçları verilmiştir. Bu sonuçlar Çizelge 6.8'de bulunmaktadır. Coğrafi konum değiştiği için uzaklık metriklerini direkt olarak karşılaştırmak mümkün değildir. Yalnızca ön bilgi amaçlı koyulmuştur. Doğruluk@161 metriği $D@161$, Ortalama değer O, Medyan ise M olarak gösterilmiştir. Bu sonuçlar İngilizce veri kümelerinde metin tabanlı sistemlerde en başarılı sistemlerin performanslarını göstermektedir. Belirtilmeyen sonuçlar boş bırakılmıştır.

Çizelge 6.8: Literatürde İngilizce için raporlanan performans değerleri.

	GeoText			**Twitter-US**			**Twitter-World**		
	$D@161$	O	M	$D@161$	O	M	$D@161$	O	M
[64]	38	844	389	54	554	120	34	1456	415
[84]	-	-	-	48	686	191	31	1669	509
[10]	-	581	425	-	-	-	-	-	-

Çizelge 6.9 temel karşılaştırma sonuçlarını göstermektedir. Buradaki sonuçlar, her kullanıcının 100 tweeti kullanılarak alınmıştır. GCN, MLP ve BERT modelleri karşılaştırılmıştır. Ortalama ve medyan metriklerinde en düşük olanlar en iyi sonucu vermektedir. Doğruluk@161 metriğinde ise sonuç arttıkça başarı artmaktadır. Tüm metriklere göre en iyi sonucu MLP vermektedir. Ayrıca temel karşılaştırma için rastgele atama yöntemiyle sonuçlar alınmıştır. En alt satırda bulunan sonuçlar 81 il için rastgele atama sonuçlarıdır. Rastgele sonuçlar için deney 3 defa tekrarlanmış ve ortalaması alınmıştır.

Çizelge 6.9: Temel karşılaştırma.

Yöntem	Ortalama	Medyan	Doğruluk@161
GCN	395	342	31
MLP	**349**	**296**	**36**
BERT	377	324	33
Rastgele Atama	563	510	8

Diğer bir deney olarak kullanıcı başına alınan tweet sayısı incelenmiştir. Tweet sayısının etkisi Çizelge 6.10'de bulunmaktadır. İlk deneyde her kullanıcının 100 tweeti kullanılırken, burada sırayla her algoritma için 75, 50 ve 25 tweet kullanılmıştır. Bu sonuçlara göre, tweet sayısı düştükçe performans azalmaktadır. Yani bir kullanıcı ne kadar çok tweet attıysa konum tespiti yapmak o kadar kolaylaşmaktadır. Ayrıca veri sayısı azaldığında BERT modelinin MLP modelinden daha başarılı sonuç verdiği görülmektedir. Kullanılan önceden eğitilmiş BERT modelinin token sınırı 512 olduğu için kullanılan tweet sayısı en fazla 100 olarak seçilmiştir. Kullanıcı tweet sayısı sırasıyla 25, 50, 75, 100 olarak seçildiğinde ortalama token uzunluğu sırasıyla 269, 528, 776, 1015 olmaktadır. Ayrıca 512 tokendan uzun metinler modelde kesilerek kullanılmaktadır. 25, 50, 75, 100 tweetli kullanıcıların token kesilme oranı sırasıyla %4, %45, %70, %87 olmuştur. Bu sebeple tweet sayısı belli bir seviyenin üzerine çıktığında kullanılan metin uzunluğu sabit kalacağı için başarım sabitlenecektir.

Çizelge 6.10: Tweet sayısının etkisi.

Tweet Sayısı	Yöntem	Ortalama	Medyan	Doğruluk@161
75	GCN	406	360	27
	MLP	396	344	30
	BERT	**366**	**314**	**34**
50	GCN	419	365	26
	MLP	402	353	29
	BERT	**372**	**317**	**33**
25	GCN	446	289	22
	MLP	429	388	26
	BERT	**385**	**321**	**31**

Konum tespitinde yapılan diğer bir deney olan veri seçme deney sonuçları Çizelge 6.11'da bulunmaktadır. Bu deneyde kullanıcıların tüm metinleri değil, yalnızca seçilen metinleri kullanılmıştır. Sadece il ve ilçe isimleri bulunan ve kelime temsillerinde yakın kelimeler bulunan metinler alınmıştır. Bu sonuçlarda en iyi modelin MLP olduğu görülmektedir. Fakat veri sayısı yine azaldığı için, önceki deneylerde alınan sonuçlardan düşük sonuç çıkmaktadır.

Çizelge 6.11: Veri seçme deneyi.

Yöntem	Ortalama	Medyan	Doğruluk@161
GCN	401	368	19
MLP	**382**	**341**	**22**
BERT	390	349	21

7. SONUÇ VE TARTIŞMA

Sonuç olarak, dönüştürücü dil modellerine etkisi hassas ayar yapmak için çeşitli veri mühendisliği yöntemleri üç farklı görev için incelenmiştir. Kontrole değer iddiaların tespiti görevi Türkçe, İngilizce, Arapça, İspanyolca ve Bulgarca dilleri kullanılarak detaylıca araştırılmıştır. Taraf tespiti için Türkçe ve İngilizce dilleri kullanılmıştır ve domain olarak futbol, siyaset, ekonomi ve sağlık gibi çeşitli konular incelenmiştir. Coğrafi konum tespiti ise Türkçe dilinde gerçekleştirilmiştir. Görevlerde kullanılan veri mühendisliği yöntemleri karşılaştırılmış ve her göreve özel başarılı sonuçlar veren yöntemler kıyaslanmıştır. Bu yöntemler göreve özel sonuçlar olsa da, yöntemlerin tüm sınıflandırma görevlerinde kullanılabileceği değerlendirilmektedir. Önceden eğitilmiş dönüştürücü dil modelleri birçok görevde en yüksek performansı gösterse de, hala farklı yöntemlerle en iyi sonuç alınan görevler bulunmaktadır. Bu sebeple coğrafi konum tespiti görevinde BERT modelinin yanında en yüksek performans gösteren GCN ve MLP algoritmaları da karşılaştırılmıştır.

Kontrole değer iddiaların tespiti görevi için, dile özel eğitim, veri dağılımı eşitleme, makine çevirisi, zayıf denetim ve çapraz dilli eğitim yöntemleri kullanılmıştır. Türkçe, İngilizce, Arapça, İspanyolca ve Bulgarca dilleri için her bir yöntem denenmiş ve her dilde en başarılı çıkan sonuç raporlanmıştır. Bu sonuçlara göre başarılı bir hassas ayar yapmak için dile özel önceden eğitilmiş dönüştürücü dil modeli kullanmak gereklidir. Aynı zamanda veri dağılımını eşitlemek performansı arttırmaktadır. Metinlere ön işleme yapmanın her zaman sonucu iyileştirilmediği görülmüştür. Ayrıca bu sonuçlara göre CTL21 labında Türkçe ve İspanyolca dillerinde en iyi performans gösteren model sunulmuştur.

Taraf tespiti görevinde, eğitim verisini değiştirme amacıyla domain bilgisi, tarafın hedefi, kullanılan dil ve veri boyutu deneyleri yapılmıştır. Türkçe ve İngilizce dilleri kullanılmıştır. Buradaki deney sonuçlarına göre, tek hedefli taraf tespiti yapmak yerine çok hedefli taraf tespiti yapmak performansı arttırmaktadır. Hedef sayısı arttığında veri boyutunun artmasının bu sonuçta etkisi olmuştur. Bununla birlikte, taraf tespitinin başarılı çalışması için bir domaine özel yapılması gerekmektedir. Domain dışı tahminler yapıldığında başarılı sonuçlar alınmamıştır. Çapraz domain deneylerine göre, yalnızca kendi domaininde farklı hedefler için hazırlanmış veriler de olsa başarılı çalışmaktadır. Çapraz dilli deneylerde çok dilli önceden eğitilmiş dönüştürücü dil modelleri kullanmak, az miktarda kendi eğitim dili verisi kullanmaya göre daha düşük skor vermektedir. İki farklı dilin birlikte kullanılması yalnızca yabancı dil kullanmaya göre daha iyi sonuçlar vermiştir. Buna rağmen farklı dillerdeki veriyi birleştirmek, yalnızca kendi dilinde eğitim verisiyle model eğitmenin başarısını yakalayamamıştır.

Coğrafi konum tespiti görevinde, eğitim verisini değiştirme amacıyla veri kümesi büyüklüğü, konuya özel veri filtreleme ve zayıf deneyim deneyleri Türkçe dili kullanılarak yapılmıştır. Bu deney sonuçlarına göre, kullanıcı başına düşen veri boyutu arttıkça başarı artmaktadır. Konuya özel filtreleme yaparak veriyi daha az ama daha kaliteli hale getirmek ve zayıf denetim kullanmak başarıyı düşürmüştür. Coğrafi konum tespiti için metin tabanlı çözümlerde kullanıcı paylaşımını en üst düzeyde tutmak performansı arttırmaktadır. Fakat belli bir paylaşım sayısının üzerinde, dönüştürücü dil modellerindeki uzunluk sınırından dolayı, metin tabanlı çalışan MLP algoritması daha iyi sonuç vermektedir. Metin verisi azaldığında ise BERT modeli en iyi performansı vermektedir.

Üzerinde çalışılan problemlerin tümü bir sınıflandırma problemidir. Üç farklı görevde bu deneyler yapıldığı için, çalışmadan çıkarılan hassas ayar yöntemlerinin, sınıflandırma görevlerinin tamamında kullanılacağı değerlendirilmektedir.

Yapılan tüm deneylerin sonucunda, dile özel önceden eğitilmiş dil modellerinin çok dilli modellere göre daha başarılı sonuçlar verdiği görülmüştür. Ayrıca hassas ayar yaparken etiketli verinin domain özelinde ve kaliteli olması performansı arttırmaktadır. Ayrıca alt örnekleme ile veya diğer veri arttırma yöntemleri ile veri dağılımını eşitlemek başarıyı arttırmaktadır. Veri arttırma yöntemi olarak aktif öğrenme, makine çevirisi, zayıf denetim, otomatik etiketleme gibi yöntemler kullanmak mümkündür. Farklı dildeki etiketli verileri kullanarak da eğitim verisi arttırılabilir. Fakat bu veriyle çok dilli hassas ayar yapmak, tek dilli eğitim verisine göre yeterli performansı göstermemiştir.

Bu çalışmada uygulanan yöntemler, doğru bir karşılaştırma için aynı parametreler kullanılarak yapılmıştır. Sonraki çalışmalarda, etkili hassas ayar yapmak için veri mühendisliği yöntemlerinin yanında, parametre optimizasyonlarının incelenmesi planlanmaktadır. Bununla birlikte, bu çalışmada incelenen doğal dil işleme görevleri temel olarak metin sınıflandırma çatısı altında kalmaktadır. Sonraki çalışmalarda görev sayısı arttırılarak yalnızca sınıflandırma için değil, diğer doğal dil işleme görevlerine de etkili hassas ayar yapma yöntemleri incelenecektir.

KAYNAKLAR

[1] **Agez, R., Bosc, C., Lespagnol, C., Mothe, J., and Petitcol, N.** (2018). IRIT at CheckThat! 2018.

[2] **Alkhalifa, R., Yoong, T., Kochkina, E., Zubiaga, A., and Liakata, M.** (2020). qmul-sds at checkthat! 2020: determining covid-19 tweet check-worthiness using an enhanced ct-bert with numeric expressions.

[3] **Allaway, E., and McKeown, K.** (2020). zero-shot stance detection: A dataset and model using generalized topic representations.

[4] **Antoun, W., Baly, F., and Hajj, H.** (2021). AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

[5] **Atanasova, P., Nakov, P., Karadzhov, G., Mohtarami, M., and Da San Martino, G.** (2019). overview of the clef-2019 checkthat! lab on automatic identification and verification of claims. task 1: Check-worthiness.

[6] **Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A., Shaar, S., and Ali, Z.** (2020). Overview of CheckThat! 2020 automatic identification and verification of claims in social media. In *Proceedings of the 11th International Conference of the CLEF Association.*

[7] **Barrón-Cedeño, A., Elsayed, T., Nakov, P., Martino, G. D. S., Hasanain, M., Suwaileh, R., and Haouari, F.** (2020). checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media. *Advances in Information Retrieval*.

[8] **Ben Zaken, E., Ravfogel, S., and Goldberg, Y.** (2021). bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models.

[9] **Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J.** (2020). spanish pre-trained bert model and evaluation data.

[10] **Cha, M., Gwon, Y., and Kung, H.** (2015). twitter geolocation and regional classification via sparse coding. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[11] **Cheema, G. S., Hakimov, S., and Ewerth, R.** (2020). check_square at checkthat! 2020: Claim detection in social media via fusion of transformer and syntactic features.

[12] **Chérif, S.** (2018). stance detection in tweets using a majority vote classifier.

[13] **Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V.** (2019). unsupervised cross-lingual representation learning at scale.

[14] **Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V.** (2020). unsupervised cross-lingual representation learning at scale.

[15] **Dai, A. M., and Le, Q. V.** (2015). semi-supervised sequence learning. *Advances in neural information processing systems*.

[16] **Darwish, K., Stefanov, P., Aupetit, M., and Nakov, P.** (2020). unsupervised user stance detection on twitter.

[17] **Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.** (2018). bert: Pre-training of deep bidirectional transformers for language understanding.

[18] **Ebrahimi, J., Dou, D., and Lowd, D.** (2016). weakly supervised tweet stance classification by relational bootstrapping. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

[19] **Eisenschlos, J. M., Ruder, S., Czapla, P., Kardas, M., Gugger, S., and Howard, J.** (2019). multifit: Efficient multi-lingual language model fine-tuning.

[20] **Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E.** (2010). a latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*.

[21] **Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., and Hovy, E.** (2021). a survey of data augmentation approaches for nlp.

[22] **Flatow, D., Naaman, M., Xie, K. E., Volkovich, Y., and Kanza, Y.** (2015). on the accuracy of hyper-local geotagging of social media content. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*.

[23] **Gardner, M. W., and Dorling, S.** (1998). artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*.

[24] **Gencheva, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., and Koychev, I.** (2017). a context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*.

[25] **Ghanem, B., Montes-y Gómez, M., Rangel, F., and Rosso, P.** (2018). UPV-INAOE-Autoritas - Check That: Preliminary approach for checking worthiness of claims.

[26] **Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A.** (2020). don't stop pretraining: adapt language models to domains and tasks.

[27] **Han, B., Cook, P., and Baldwin, T.** (2012). geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*.

[28] **Han, B., Cook, P., and Baldwin, T.** (2014). text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*.

[29] **Hansen, C., Hansen, C., Simonsen, J., and Lioma, C.** (2018). the Copenhagen team participation in the check-worthiness task of the competition of automatic identification and verification of claims in political debates of the CLEF-2018 fact checking lab.

[30] **Hansen, C., Hansen, C., Simonsen, J., and Lioma, C.** (2019). neural weakly supervised fact check-worthiness detection with contrastive sampling-based ranking loss.

[31] **Hasan, K. S., and Ng, V.** (2013). stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*.

[32] **Hasanain, M., and Elsayed, T.** (2020). bigIR at CheckThat! 2020: Multilingual BERT for ranking Arabic tweets by check-worthiness.

[33] **Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A. K., Sable, V., Li, C., and Tremayne, M.** (2017). claimbuster: The first-ever end-to-end fact-checking system.

[34] **Howard, J., and Ruder, S.** (2018). universal language model fine-tuning for text classification.

[35] **Huang, B., and Carley, K. M.** (2019). a large-scale empirical study of geotagging behavior on twitter. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

[36] **Hussein, A., Hussein, A., Ghneim, N., and Joukhadar, A.** (2020). damascus-team at checkthat! 2020: Check worthiness on twitter with hybrid cnn and rnn models. In *CLEF (Working Notes)*.

[37] **Jain, A., Ruohe, A., Grönroos, S.-A., and Kurimo, M.** (2020). finnish language modeling with deep transformer models.

[38] **Jaradat, I., Gencheva, P., Barrón-Cedeño, A., Màrquez, L., and Nakov, P.** (2018). claimrank: Detecting check-worthy claims in arabic and english. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*.

[39] **Kartal, Y. S., and Kutlu, M.** (2020). TOBB ETU at CheckThat! 2020: Prioritizing English and Arabic claims based on check-worthiness.

[40] **Kipf, T. N., and Welling, M.** (2016). semi-supervised classification with graph convolutional networks.

[41] **Küçük, D.** (2017). stance detection in turkish tweets.

[42] **Küçük, D., and Can, F.** (2019). a tweet dataset annotated for named entity recognition and stance detection.

[43] **Lai, M., Cignarella, A., Hernandez Farias, D., Bosco, C., Patti, V., and Rosso, P.** (2020). multilingual stance detection in social media political debates.

[44] **Lample, G., and Conneau, A.** (2019). cross-lingual language model pretraining.

[45] **Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J.** (2020). biobert: a pre-trained biomedical language representation model for biomedical text mining.

[46] **Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V.** (2019). roberta: A robustly optimized bert pretraining approach.

[47] **Lourentzou, I., Morales, A., and Zhai, C.** (2017). text-based geolocation prediction of social media users with neural networks. In *2017 IEEE International Conference on Big Data (Big Data)*.

[48] **Maiya, A. S.** (2020). ktrain: A low-code library for augmented machine learning.

[49] **Martinez-Rico, J. R., Araujo, L., and Martinez-Romo, J.** (2020). nlp&ir@ uned at checkthat! 2020: A preliminary approach for check-worthiness and claim retrieval tasks using neural networks and graphs. In *CLEF (Working Notes)*.

[50] **Martinez-Rico, J. R., Martinez-Romo, J., and Araujo, L.** (2021). l.: Nlp&ir@ uned at checkthat! 2021: check-worthiness estimation and fake news detection using transformer models.

[51] **Mass, Y., and Roitman, H.** (2020). ad-hoc document retrieval using weak-supervision with bert and gpt2. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[52] **McDonald, T., Dong, Z., Zhang, Y., Hampson, R., Young, J., Cao, Q., Leidner, J. L., and Stevenson, M.** (2020). the university of sheffield at checkthat! 2020: Claim identification and verification on twitter. In *CLEF (Working Notes)*.

[53] **Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C.** (2016). SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*.

[54] **Mohammad, S. M., Sobhani, P., and Kiritchenko, S.** (2017). stance and sentiment in tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*.

[55] **Moreo, A., Esuli, A., and Sebastiani, F.** (2016). distributional random oversampling for imbalanced text classification. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*.

[56] **Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeno, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Babulkov, N., et al.** (2021). the clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *European Conference on Information Retrieval*.

[57] **Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeño, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Mansour, W., et al.** (2021). overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news.

[58] **Nikolov, A., Da San Martino, G., Koychev, I., and Nakov, P.** (2020). Team_Alex at CheckThat! 2020: Identifying check-worthy tweets with transformer models.

[59] **Patwari, A., Goldwasser, D., and Bagchi, S.** (2017). tathya: A multi-classifier system for detecting check-worthy statements in political debates. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*.

[60] **Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L.** (2018). deep contextualized word representations.

[61] **Phang, J., Févry, T., and Bowman, S. R.** (2018). sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks.

[62] **Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I.** (2018). improving language understanding by generative pre-training.

[63] **Radiya-Dixit, E., and Wang, X.** (2020). how fine can fine-tuning be? learning efficient language models. In *International Conference on Artificial Intelligence and Statistics*.

[64] **Rahimi, A., Cohn, T., and Baldwin, T.** (2017). a neural model for user geolocation and lexical dialectology.

[65] **Rahimi, A., Cohn, T., and Baldwin, T.** (2018). semi-supervised user geolocation via graph convolutional networks.

[66] **Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C.** (2017). snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*.

[67] **Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X.** (2021). a survey of deep active learning. *ACM Computing Surveys (CSUR)*.

[68] **Ribeiro, M. T., Singh, S., and Guestrin, C.** (2016). "why should i trust you?" explaining the predictions of any classifier.

[69] **Roller, S., Speriosu, M., Rallapalli, S., Wing, B., and Baldridge, J.** (2012). supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*.

[70] **Sabou, M., Bontcheva, K., Derczynski, L., and Scharl, A.** (2014). corpus annotation through crowdsourcing: Towards best practice guidelines.

[71] **Schlicht, I. B., de Paula, A. F. M., and Rosso, P.** (2021). upv at checkthat! 2021: Mitigating cultural differences for identifying multilingual check-worthy claims.

[72] **Schröder, C., and Niekler, A.** (2020). a survey of active learning for text classification using deep neural networks.

[73] **Schweter, S.** (2020). berturk - bert models for turkish.

[74] **Sellam, T., Das, D., and Parikh, A. P.** (2020). bleurt: Learning robust metrics for text generation.

[75] **Shleifer, S.** (2019). low resource text classification with ulmfit and backtranslation.

[76] **Suh, Y., Yu, J., Mo, J., Song, L., and Kim, C.** (2017). a comparison of oversampling methods on imbalanced topic classification of korean news articles. *Journal of Cognitive Science*.

[77] **Sun, C., Qiu, X., Xu, Y., and Huang, X.** (2019). how to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*.

[78] **Tong, S., and Koller, D.** (2001). support vector machine active learning with applications to text classification. *Journal of machine learning research*.

[79] **Varma, P., and Ré, C.** (2018). snuba: Automating weak supervision to label training data. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*.

[80] **Vasileva, S., Atanasova, P., Màrquez, L., Barrón-Cedeño, A., and Nakov, P.** (2019). it takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.

[81] **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, ., and Polosukhin, I.** (2017). attention is all you need. In *Advances in neural information processing systems*.

[82] **Williams, E., Rodrigues, P., and Novak, V.** (2020). accenture at checkthat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models.

[83] **Williams, E., Rodrigues, P., and Tran, S.** (2021). accenture at checkthat! 2021: Interesting claim identification and ranking with contextually sensitive lexical training data augmentation.

[84] **Wing, B., and Baldridge, J.** (2014). hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

[85] **Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.** (2019). huggingface's transformers: State-of-the-art natural language processing.

[86] **Xu, C., Paris, C., Nepal, S., and Sparks, R.** (2018). cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

[87] **Yang, B., Sun, J.-T., Wang, T., and Chen, Z.** (2009). effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*.

[88] **Yasser, K., Kutlu, M., and Elsayed, T.** (2018). bigir at CLEF 2018: Detection and verification of check-worthy political claims. In *Working Notes of CLEF 2018*.

[89] **Yu, T., and Joty, S.** (2021). effective fine-tuning methods for cross-lingual adaptation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

[90] **Yu, Y., Zuo, S., Jiang, H., Ren, W., Zhao, T., and Zhang, C.** (2020). fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach.

[91] **Zengin, M. S., Kartal, Y. S., and Kutlu, M.** (2021). tobb etu at checkthat! 2021: data engineering for detecting check-worthy claims.

[92] **Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., and Artzi, Y.** (2020). revisiting few-sample bert fine-tuning.

[93] **Zheng, B., Dong, L., Huang, S., Wang, W., Chi, Z., Singhal, S., Che, W., Liu, T., Song, X., and Wei, F.** (2021). consistency regularization for cross-lingual fine-tuning.

[94] **Zheng, C., Jiang, J.-Y., Zhou, Y., Young, S. D., and Wang, W.** Social media user geolocation via hybrid attention. In *(2020). Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[95] **Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., and Liu, T.-Y.** (2020). incorporating bert into neural machine translation.

[96] **Zuo, C., Karakas, A., and Banerjee, R.** (2018). a hybrid recognition system for check-worthy claims using heuristics and supervised learning.