TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İDDİALARIN TEYİT GEREKLİLİĞİNE GÖRE ÖNCELİKLENDİRİLMESİ

YÜKSEK LİSANS TEZİ

Yavuz Selim KARTAL

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Öğretim Üyesi Mücahid KUTLU

Mart 2021

ÖZET

Yüksek Lisans Tezi

İDDİALARIN TEYİT GEREKLİLİĞİNE GÖRE ÖNCELİKLENDİRİLMESİ

Yavuz Selim KARTAL

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

Tez Danışmanı: Dr. Öğretim Üyesi Mücahid KUTLU

Tarih: Mart 2021

Yanlıs¸ bilgiler, internette inanılmaz bir s¸ekilde her gün yayılmaktadır ve toplumlar üzerindeki olumsuz etkileri tehlikeli seviyelere ulas¸mıs¸tır. Yanlıs¸ bilgilerin en önemli düşmanı doğruluk kontrolü yapanlardır. Ancak yanlış bilgilerin yayılma hızı göz önüne alındığında, doğruluk kontrolü yapmak yavaş olduğundan tüm iddiaların kontrol edilmesi mümkün olmamaktadır. Bu yüzden, iddiaları teyit gerekliliklerine göre önceliklendirerek doğruluk kontrolü yapanlara yardımcı olacak sistemlerin gelis¸tirilmesi ve bu konuda farkındalık olus¸turulması büyük önem tas¸ımaktadır. Bu alandaki bir diğer problem ise, geliştirilecek sistemler için kullanılabilecek veri kaynaklarının çoğunlukla İngilizce olmak üzere sınırlı olmasıdır. Bu tez çalıs¸masında öncelikle Türkçe için ilk teyit gerektiren iddia veri kümesi olan TrClaim-19 hazırlanmıs¸tır. TrClaim-19, 2287 tane etiketli tweet içermenin yanı sıra, teyit gerektirme özelliklerinin daha iyi anlaşılmasını sağlayacak olan teyit gerektirme gerekçeleri de sunulmus¸tur. Bu gerekçeler, iddiaların konularının ve muhtemel negatif etkilerinin teyit gerektirmeye sebep olan ana etkenler olduğunu öne sürmektedir. Tez çalıs¸masında ayrıca, iddiaları teyit gerekliliklerine göre önceliklendirmek için BERT modelinin ve çeşitli özniteliklerin kullanıldığı karma bir model de önerilmis¸tir. Kullanılan öznitelikler, yerel bölgeye özgü tartıs¸malı konular, kelime vektörleri, POS etiketleri ve daha fazlasını içermektedir. Buna ek

olarak, teyit gerektiren verileri artırma, aktif öğrenme ve farklı dillerde verileri kullanma gibi veri kümesi boyutunu artırmanın farklı yolları üzerine çalıs¸malar yapılmıs¸tır. Kapsamlı deneyler sonucunda, modelimizin, CLEF Check That! Lab 2018 and 2019 test koleksiyonlarındaki en iyi modellerden daha başarılı olduğu gözlemlenmiştir. Modelimiz, eğitim verilerindeki teyit gerektiren örnekler artırıldığında, Check That! Lab 2020'in test koleksiyonu için de şimdiye kadar bildirilen en iyi MAP puanını elde etmiştir. Çok dilli eğitimin ise Arapça ve Türkçe iddiaları önceliklendirmek için etkili olduğu, ancak bunun İngilizce için geçerli olmadığını gözlemlenmis¸tir.

Anahtar Kelimeler: Teyit Gerektiren İddialar, Doğruluk Kontrolü, Yanlış Bilgi.

ABSTRACT

Master of Science

PRIORITIZING CHECK-WORTHY CLAIMS

Yavuz Selim KARTAL

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

Supervisor: Dr. Öğretim Üyesi Mücahid KUTLU

Date: March 2021

The massive amount of misinformation spreading on the Internet on a daily basis has enormous negative impacts on societies. In order to combat against misinformation and its negative outcomes, fact-checking websites detect the veracity of claims . However, fact-checking is an extremely time-consuming process and human fact-checkers are not able detect the veracity of all claims spread on the Internet. Therefore, we need systems to help fact-checkers in the combat against misinformation and to raise public awareness of this important problem. Anot- her problem is that available data resources to develop effective systems are li- mited and the vast majority of them is for English. In this thesis, we introduce TrClaim-19, which is the very first labeled dataset for Turkish check-worthy cla- ims. TrClaim-19 consists of labeled 2287 Turkish tweets with annotator rationa- les, enabling us to better understand the characteristics of check-worthy claims. The rationales we collected suggest that claims' topics and their possible negative impacts are the main factors affecting their check-worthiness. In this thesis, we also propose a hybrid model which combines BERT model with various featu- res to prioritize claims based on their check-worthiness. Features we use include domain-specific controversial topics, word embeddings, POS tags, and others. In addition, we explore various ways of increasing labeled data size to effectively

train the models such as increasing positive samples, active learning, and utilizing labeled data in other languages. In our extensive experiments, we show that our model outperforms all state-of-the-art models in test collections of CLEF Check That! Lab 2018 and 2019. In addition, when positive samples are increased in the training set, our model achieves the best MAP score reported so far for the test collection of Check That! Lab 2020. Furthermore, we show that cross-lingual training is effective for prioritizing Arabic and Turkish claims, but not for English.

Keywords: Check-Worthy Claims, Fact-Checking, Misinformation.

TEŞEKKÜR

Yüksek lisans sürecim boyunca beni her zaman destekleyen, değerli yardım ve katkılarıyla beni yönlendiren, sürekli yaptığımız çalışmalarla bu sürecin en verimli şekilde geçmesini sağlayan ve danışmanlığında çalışmalarını tamamlayan ilk öğrencisi olmanın gururunu duyduğum hocam Dr. Öğretim Üyesi Mücahid KUTLU, bu tezi değerlendiren ve kıymetli görüşlerini paylaşan Prof. Dr. Fazlı CAN ve Prof. Dr. Osman ABUL ve tecrübelerinden faydalandığım TOBB Ekonomi ve Teknoloji Üniversitesi Bilgisayar Mühendisliği Bölümü öğretim üyelerine çok teşekkür ederim.

Bu süreçte her zaman yanımda olan ve yardımlarını esirgemeyen başta eşim Emine ve kız kardeşim Sevim olmak üzere aileme teşekkür etmek isterim. Yaptıkları katkılardan dolayı Yasin Furkan Aktaş, Nuri Altın, Caner Sicimali, Büşra Güvenen, Nursena Ünlü ve Zeynep Memet'e ayrı ayrı teşekkür ederim.

İÇİNDEKİLER

Sayfa

ŞEKİL LİSTESİ

ÇI˙ZELGE LI˙STESI˙

KISALTMALAR

MAP : Ortalama Hassasiyet (Mean Average Precision)
BERT : Çift Yönlü Dönüs¸türücü Kodlayıcısı (Bidirectional Encoder for Transformers)
POS : Konus¸ma Bölümü (Part-of-Speech)
LR : Lojistik Regresyon (Logistic Regression)
SVM : Vektör Destek Makinesi (Support Vector Machine)
kNN : En Yakın k Koms¸u (k Nearest Neighbor)
FFNN : İleri Beslemeli Sinir Ağları (Feed Forward Neural Network)
RNN : Özyineli Sinir Ağları (Recurrent Neural Network)
MLP : Çok Katmanlı Algılayıcı (Multi Layer Perceptron) **RF** : Rassal Ormanlar (Random Forest)
TF-IDF : Terim Sıklığı-Ters Doküman Sıklığı (Term Frequency-Inverted Document Frequency)
BoW : Kelime Torbası (Bag-of-Words)
LDA : Gizli Semantik Analiz (Latent Semantic Analysis)
LSTM : Uzun Kısa Vadeli Bellek (Long Short Term Memory)
RP : R Kesinlik (R Precision)
P@k : İlk K Sonuçta Kesinlik (Precision at k)

1.G İRİŞ

2013 yılında, Dünya Ekonomik Forumu (DEF) tarafından hazırlanan rapora göre dijital yanlıs¸ bilgi, on yılın en büyük küresel risklerden biri olarak sıralanmıs¸- tır[1]. İnternette yayılan yanlış bilgilerin sebep olduğu birçok olay, WEF'in öngörüsünü desteklemektedir. "Pizzagate" yalan haberleri yüzünden ortaya çıkan çatışma[2], borsa hisselerindeki kayıplar [3] ve aşılamaya karşı güvensizliğin artması [4], yanlıs¸ bilgilerin yayılmasına örnek olarak gösterilebilir. COVID-19 pandemisinin başlangıcından bu yana da, gerçek bilginin değerini ve sağlık sorunları hakkında yanlış bilginin nasıl ölümcül olabileceğini gösteren birçok olay gözlemlenmiştir. Buna örnek olarak, koronavirüs bulaşmasını engellemek için sağlığa zararlı kimyasalların kullanılması verilebilir[5].

Yanlıs¸ bilgi, uluslararası olarak da yayılmaktadır. Benzer yanlıs¸ bilgi yayılımının 2019 Avrupa seçimlerinde, Avrupa ülkelerinde farklı dillerde gerçekleştiği gözlemlenmis¸tir [13]. Morstatter vd. [29] de ˙Ingilizce konus¸ulan Twittersphere'deki birçok hesabın 2017 Almanya Federal seçimlerini etkilemeye çalıştığını belirtmişlerdir. Bu nedenle, dijital yanlıs¸ bilgiye etkili bir çözüm için birçok ülke ve dile odaklanılması gerekmektedir.

[1]http://reports.weforum.org/global-risks-2013

[2]www.nytimes.com/2016/12/05/business/media/comet-ping-pong-pizza-shooting-fake-news-consequences.html

[3]www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market

[4]www.washingtonpost.com/news/wonk/wp/2014/10/13/the-inevitable-rise-of-ebola-conspiracy-theories

[5]https://www.reuters.com/article/us-health-coronavirus-disinfectants-idUSKBN23C2P2

Yanlış bilgiye ve olumsuz sonuçlarına karşı mücadele etmek için, doğruluk kontrolü yapan Snopes[6] ve Teyit[7] gibi web siteleri mevcuttur. Bu siteler, internet üzerinde yayılan iddiaların doğruluğunu kontrol ederek bulgularını okurlarıyla paylaşmaktadırlar [7]. Bununla birlikte, iddia doğruluk kontrolü, son derece zaman alan bir süreçtir ve sadece bir iddianın kontrol edilmesi yaklas¸ık bir gün sürebilmektedir [21]. Gazetecilerin bu önemli çabaları yanlıs¸ bilgilerin yayılmasını azaltmaya yardımcı olurken, Vosoughi vd. [37] yanlıs¸ haberlerin gerçek haberden sekiz kat daha hızlı yayıldığını belirtmiştir. Bu nedenle, yanlış bilgiyle mücadelede doğruluk kontrolü yapanlara yardımcı olan sistemlere olan ihtiyaç büyük önem tas¸ımaktadır.

Nakov vd. [30], bir doğruluk kontrol sisteminin ilk görevinin, bir ifadenin teyit gerektiren bir iddia içerip içermediğini tespit etmek olduğunu belirtmişlerdir. Sosyal medya platformlarında paylaşılan iletilerin boyutu düşünüldüğünde, insan doğruluk kontrolcülerinin internette yayılan tüm iddiaların doğruluğunu tespit edemediklerinden, değerli zamanlarını en önemli iddiaların gerçekliğini kontrol etmek için harcaması önemlidir. Bu nedenle, sosyal medya paylas¸ımlarını, haber makalelerini ve politikacıların açıklamalarını izleyerek teyit gerektiren iddiaları önceliklendiren otomatik sistemlere ihtiyaç vardır. Ayrıca, internet kullanıcıları yanlıs¸ bilginin olası olumsuz sonuçlarının farkında olmadan ve gördükleri her şeyi doğruluğunu sorgulamadan paylaşmaya devam ederlerse, mükemmel bir doğruluk kontrol sistemi bile yanlıs¸ bilgiyle mücadelede etkili olmayacaktır. Bu nedenle, kullanıcıları sosyal medya platformlarında doğruluğu teyit gerektiren bir ileti yayınlamadan önce uyaran araçların, kullanıcıların bu iletileri paylas¸mayı yeniden düşünmelerinde ve yanlış bilgilerin yayılmasını azaltmada etkili olabileceği düs¸ünülmektedir.

Teyit gerektiren iddiaların önceliklendirilmesi için gelis¸tirilen modeller mevcuttur [21, 22, 31]. Ayrıca, Conference and Labs of Evaluation Forum (CLEF), 2018'den beri bu konuda CheckThat! Lab (CTL) adında bir paylas¸ımlı görev organize edilmektedir [3, 6, 30]. Önceki çalışmaların çoğu İngilizce iddialara odaklanırken [21, 22, 31] Arapça için de çalıs¸malar mevcuttur [4, 5].

[6]https://www.snopes.com/
[7]https://teyit.org/

Çizelge 1.0.1: CTL'de Sunulan Veri Kümesinden Bir Bölüm. Teyit gerektiren iddia **koyu** yazılmıs¸tır.

CLINTON:	Temiz enerji kullanın.
CLINTON:	Bazı ülkeler 21. yüzyılın temiz enerji süper gücü olacak.
CLINTON:	**Donald, iklim değişikliğinin Çinliler tarafından yapılan bir aldatmaca olduğunu düşünüyor.**
CLINTON:	Bunun gerçek olduğunu düşünüyorum.

Teyit gerektiren iddialar için Çizelge 1.0.1'de örnek gösterilmis¸tir. CTL veri kümesinden alınan bu örnek, Amerika'da seçim öncesi yapılan tartıs¸ma programından alıntıdır. Örnekte, Hillary Clinton'ın Donald Trump hakkında bir iddia öne sürmüştür. İddiada bulunan "iklim değişikliğinin Çinliler tarafından yapılan bir aldatmaca" ifadesi ise doğruluğu kontrol etmeye değerdir, bu yüzden bu iddia teyit gerektirmektedir.

Bu tez çalıs¸masında, öncelikle teyit gerektiren iddialar üzerine ilk Türkçe veri kümesi (TrClaim-19) olus¸turularak veri kümesi üzerinde bu konuda yapılacak çalıs¸malara yardımcı olacak analizler yapılmıs¸tır. Bunun yanı sıra, teyit gerektiren iddiaları önceliklendirmek için BERT tabanlı karma bir sistem önerilerek CTL'18 ve CTL'19 veri setlerinde en bas¸arılı sonuçlar elde edilmis¸tir. Ayrıca, çes¸itli yöntemler kullanılarak veri kümesinin genişletmenin performansı nasıl etkileyeceği üzerine çalıs¸malar yapılmıs¸tır. Bu çalıs¸malar sonucunda, teyit gerektiren iddiaların önceliklendirilmesi için etiketli veri miktarını rastgele artırmanın performansı her zaman artırmadığı, aktif öğrenme kullanılarak veri artırmanın biraz daha başarılı olduğu ve sadece teyit gerektiren iddiaların sayısını artırmanın testlerde kullanılan iki veri kümesinde en başarılı sonuçlara ulaştığı gözlemlenmiştir. Tezin bir diğer araştırmasında ise, farklı dillerden veriler kullanarak veri kümeleri genişletildiğinde İngilizce iddialara diğer dillerden veriler eklendiğinde daha başarısız, Türkçe ve Arapça verilere eklendiğinde ise daha başarılı sonuçlar elde edildiği gözlemlenmis¸tir.

Yaptığımız çalışmalar sonucunda elde etmiş olduğumuz çıkarımlar ile literatüre katkı sağladığımız konular şunlardır:

- Türkçe için ilk etiketli teyit gereklilik veri kümesi olan TrClaim-19 olus¸turulmus¸ ve paylas¸ılmıs¸tır[8].
- TrClaim-19, aynı zamanda, etiketleme gerekçeleri sunan ilk teyit gereklilik veri kümesidir.
- Teyit gereklili ğinin öznelliği, özellikle uzman ve uzman olmayan doğruluk kontrolcülerinin kararları üzerinden incelenmis¸tir.
- Gelecek çalıs¸malara yardımcı olmak için, TrClaim-19 kullanılarak 4 farklı referans modelin değerlendirmesi yapılmıştır.
- Teyit gereklilik görevi için çes¸itli öznitelikler incelenerek etkileri raporlanmıs¸tır.
- Teyit gereklilik görevi için öznitelik mühendisli ği ve dönüştürücü modeli kullanan bir model önerilmis¸tir.
- Dönüs¸türücü modelinin ve önerilen modelimizin performansını artırmak için veri kümesinin en iyi nasıl artırılması gerektiği incelenmiştir.
- Çok dilli e ğitimin, İngilizce olmayan iddiaları önceliklendirmede etkili olduğu gözlemlenmiştir.
- Üç CTL koleksiyonu üzerinde yapılan deneylerde, modelimiz CTL'18 ve CTL'19 için gelis¸tirilen en iyi modellerden daha bas¸arılı olmus¸tur. Modelimiz, eğitim veri kümesi genişletildiğinde, CTL'20-ED koleksiyonunda da en bas¸arılı sonuçları alde etmis¸tir.
- Modelimiz sonuçları ve veri kümeleri üzerine nitel analizler sunulmus¸tur.

[8]https://github.com/YSKartal/TrClaim19

Bu tezin organizasyonu s¸us¸ekildedir: Bölüm 2 teyit gereklilik üzerine yapılmıs¸ literatür araştırmalarını, Bölüm 3 Türkçe için hazırlandığımız ilk teyit gereklilik veri kümesinin detaylarını, Bölüm 4 teyit gerektiren iddiaların önceliklendirilmesi için önerdiğimiz modelin detaylarını, deneylerini ve nitel analizini, Bölüm 5 eğitim veri kümesini genis¸letmek için kullanılan yöntemleri ve deneylerini, Bölüm 6 ise tez sonucunda elde edilmis¸ çıkarımları içermektedir.

2.L İTERATÜR ARAŞTIRMASI

2016'da yapılan ABD başkanlık seçimleri, doğruluk kontrol çalışmalarını hızlandıran ana etkenlerden biri olarak kabul edilmektedir. Bu nedenle, teyit gerektiren iddia tespiti üzerine yapılan önceki çalışmalarda, veri kümeleri olarak çoğunlukla ABD'li politikacıların tartıs¸ma ve konus¸ma metinleri kullanılmıs¸tır [21, 25]. Bu yüzden, mevcut çalışmaların çoğu İngilizce üzerine olmakla beraber, Arapça [22] gibi farklı diller üzerine yapılan çalıs¸malar az da olsa bulunmaktadır.

ClaimBuster [21], bu konuyla ilgili yapılan ilk çalıs¸malardandır. ClaimBuster'da TF-IDF, adlandırılmıs¸ varlıklar, POS etiketleri ve duygu gibi birçok öznitellik kullanılmıs¸tır. Patwari vd. [31] ise, varlıkların geçmis¸i, BoW, konu bas¸lıkları ve POS etiketleri gibi çes¸itli öznitellikler üzerinde çalıs¸ma yapmıs¸lardır. Konu bas¸lıklarını tespit etmek için, 1976'dan 2016'ya kadar tüm bas¸kanlık tartıs¸malarının yazılı metinleri üzerinde eğitilen bir LDA modelini kullanmışlardır.

Gencheva vd. [16], konu başlığı, zaman kipi, cümle uzunluğu, karşıt ifade, adlandırılmış varlık, duygu, POS etiketi ve kelime vektörü gibi çeşitli bağlamsal ve cümle seviyesinde öznitellikler kullanarak bir sinir ağı modeli geliştirmişlerdir. Jaradat vd. [22] , Gencheva vd. tarafından geis¸tirilen modeli, neredeyse aynı öznitellikleri kullanarak Arapça için uygulamıs¸lardır. Bununla birlikte, kullandıkları Arapça veriler, sadece ABD seçim tartıs¸malarının çevirisidir. Vasileva vd. [35], doğruluk kontrolü yapan dokuz saygın kuruluşu belirleyip verilen bir iddianın bu kuruluşlardan herhangi biri tarafından doğruluk kontrolü yapılıp yapılmadığını tahmin eden çok-görevli bir öğrenme modeli geliştirmişlerdir.

Check That! Labs (CTL) 2018'den beri CLEF tarafından organize edilmektedir. CTL'18'in teyit gerektiren iddiaların önceliklendirilmesi görevine yedi takım katılmıs¸tır [30]. Katılımcılar, SVM [41], kNN [17], MLP [42], RF [1] ve RNN [18] gibi çeşitli öğrenme modellerini kullanmışlardır. Kullandıkları öznitelikler arasında kelime torbası (BOW) [42], karakter n-gram [17], olumsuz ifadeler [42], adlandırılmış varlıklar [41, 42], POS etiketleri [18, 41, 42], sözdizimsel bağımlılıklar [18, 42], sözlü ifadeler [42] ve kelime vektörleri [18, 41, 42] yer almaktadır. İngilizce veri setinde, Prize de Fer ekibi [42], çes¸itli sözdizimsel ve anlamsal özelliklerle SVM-MLP kullanarak ve tartıs¸maların etkiles¸imli söylem yapısını dikkate alarak en iyi MAP puanını elde etmis¸tir.

Lespagnol vd. [25], POS etiketleri, varlıklar, kelime vektörleri ve sözdizimsel bağımlılık etiketleri ile güvenilirliği, karşıtlığı, gerçekliği, duyguları ve ifadelerin teknik özelliklerini temsil eden "bilgi besini" olarak adlandırılan çes¸itli öznitellikler kullanan lojistik regresyon, SVM ve RF gibi öğrenme modellerini incelemis¸lerdir. Kelime vektörleri ve bilgi besini özniteliklerine sahip lojistik regresyon modelleri ile CTL'18 katılımcılarından daha bas¸arılı sonuçlar elde etmis¸lerdir.

2019'da 11 ekip, CTL'19'un teyit gereklilik görevine katılmıs¸tır [3]. Katılımcılar, LSTM [10, 19], FFNN [11], SVM [34], naive bayes [8] ve lojistik regresyon [2] gibi çes¸itli modeller kullanmıs¸lardır. Öznitelik olarak ise POS etiketleri [2, 15], adlandırılmıs¸ varlıklar [2, 15], konu bas¸lıkları [2], okunabilirlik [11], duygu ifadeleri [11, 15], ve kelime vektörleri [10, 15] bas¸ta olmak üzere çes¸itli öznitelikler kullanmıs¸lardır. Kopenhag ekibi [19], her kelimenin bir kelime vektörü ve sözdizimsel bağımlılıkları ile temsil edildiği zayıf denetimli (weak supervision) bir LSTM modelini kullanarak birinci olmayı bas¸armıs¸tır. Zayıf denetim yapmak için ClaimBuster'ı kullanarak önceki tartıs¸ma metinlerini etiketlemis¸lerdir.

2020 yılında, siyasi tartıs¸ma ve konus¸ma metinlerine (5. Görev) ek olarak tweetler (1. Görev) de CTL'e dahil edilmis¸tir. 1. Görev, hem Arapça hem de İngilizce'yi kapsarken, 5. Görev yalnızca İngilizce'yi kapsayacak s¸ekilde düzenlenmis¸tir. 1. Görevde İngilizce tweetler için 12 takım model gelis¸tirmis¸ olup bunlardan sekizi BERT veya RoBERTa kullanmıs¸tır. Görevin katılımcıları fastText, adlandırılmıs¸ varlıklar ve POS etiketleri gibi çes¸itli öznitelikleri kullanarak RF, SVM, Bi-LSTM

ve CNN modellerini de incelemis¸lerdir. En bas¸arılı ekip olan Accenture [38], modellerinin as¸ırı uyumunu önlemek için ortalama havuzlama ve bırakma katmanı ile RoBERTa'yı kullanmıs¸tır. Aynı görevde, Arapça tweetler için 8 takım model gelis¸tirmis¸ ve yine en üst sıradaki üç takımın AraBERT veya MBERT kullan- dığı gözlemlenmiştir. Birinci sıradaki Accenture [38] yine RoBERTa'yı kullan- mıs¸ olup veriyi artırmak için Arapça tweet'leri otomatik olarak İngilizce'ye ve ardından tekrar Arapça'ya çevirmiştir. 5. Göreve ise 3 takım katılmıştır. Diğer görevlerin aksine, bu veri kümesinde, BERT tabanlı sistemlerin diğerlerinden daha kötü performans gösterdiği görülmüştür. NLPIR@UNED [26], glove kelime vektörleri ile Bi-LSTM modelini kullanarak birinci sırada yer almıs¸tır. Ayrıca veri kümesini genişletmeye çalıştıklarını ancak bunun performansı düşürdüğünü bildirmis¸lerdir. UAICS ekibi [27], çok sınıflı Naive Bayes ile TF-IDF vektörlerini kullanarak ikinci olurken, TOBB ETU ekibi [23] de MBERT tabanlı karma modelle üçüncü sırada yer almıs¸tır.

Bu tez çalıs¸masını mevcut çalıs¸malardan s¸u yönleriyle ayrılmaktadır. 1) Türkçe için ilk teyit gereklilik veri kümesi olan TrClaim-19 hazırlanmıs¸tır. TrClaim-19 aynı zamanda etiketleyicilerin gerekçelerini bildirdikleri ilk teyit gereklilik veri kümesidir. Bu görevin öznelliğini incelemek için her tweet üç kişi tarafından etiketlenmis¸ ve bağımsız olarak gerekçelerini bildirmişlerdir. 2) İnce ayarlanmış

BERT modeli ve diğer öznitelikleri kullanan bir model önerilmiştir. 3) İddianın konusu teyit gerekliliği etkileyebilecek bir unsur olduğundan birçok çalışmada farklı s¸ekillerde kullanılmıs¸tır [25, 31, 41]. Ancak, bu çalıs¸mada bölgesel ve gündemdeki tartışmalı konuların teyit gerekliliği etkileyebileceği düşünülmüştür. Bu yüzden, ABD gündemindeki konularla ilgili bir liste hazırlanarak öznitelik olarak kullanılmıs¸tır. Bunun yanı sıra, özel kelime listesi ve kars¸ılas¸tırma ifadelerini kullanan öznitelikler de kullanılmıştır. 4) Önceki çalışmalarda, eğitim veri kümesini artırmak için zayıf denetimleme [19], fazla örneklendirme (over sampling) [26] ve makine çevirisi (machine translition) [38] gibi yöntemler kullanılmıs¸tır. Bu çalışmada ise, hassas ayarlanmış BERT modelleri için eğitim verisinin etkileri incelenmiş ve en iyi yöntemi bulmaya çalışılmıştır. Ayrıca, çok dilli eğitim kümelerinin etkisi de incelenmiştir. Bilindiği kadarıyla, teyit gereklilik konusunda daha önceden yapılmıs¸ çok dilli çalıs¸ma bulunmamaktadır.

3 TRCLAIM-19: TÜRKÇE İÇİN TEYİT GEREKLİLİK İDDİA VERİ KÜMESİ

Araştırmacılar doğruluk kontrolüne büyük ilgi gösterirken, bu konudaki mevcut kaynaklar hala sınırlıdır ve çalışmaların büyük çoğunluğu İngilizce üzerine odaklanmış,tır. Teyit gerektiren iddiaları önceliklendirme görevi ile ilgili olarak, var olan etiketli veri kümeleri yalnızca İngilizce ve Arapça için mevcuttur [3, 30]. Bununla birlikte, DEF'in önceden belirtilen raporunda ifade edildiği gibi, yanlış bilgiler tüm ülkeleri etkileyen küresel bir sorundur.

Tezin bu bölümünde, teyit gerektiren iddiaların önceliklendirilmesi için ilk Türkçe veri kümesi hazırlanmış,tır. Fletcher vd. [13], Türkiye'deki İnternet kullanıcılarının % 49'unun haftada en az bir sahte haberle karşılaştığı belirtmişlerdir. Diğer ülkelerle kıyaslandığında bu oranın en yüksek Türkiye'de olması, Türkçe'yi doğruluk kontrolü çalış,maları için önemli hale getirmektedir. Ayrıca Altay dil ailesinin bir üyesi olan Türkçe, sondan eklemeli ve esnek sözdizimi gibi özelliklere sahip olması nedeniyle doğruluk kontrolü çalışmaları yapılan diğer dillerden ayrılmaktadır. Bu veri kümesinin hazırlanmasıyla, aras,tırma çalış,maları için yararlı bir kaynak geliştirmenin yanı sıra, aşağıdaki araştırma sorularının yanıtlarını da aranmaktadır.

- AS-1: Uzman olmayan doğruluk denetçilerinin iddiaların teyit gerekliliği konusunda anlas,ma düzeyi nedir?
- AS-2: Uzman olmayanların iddiaların teyit gerekliliği konusunda uzmanlardan farklı fikirleri var mı?

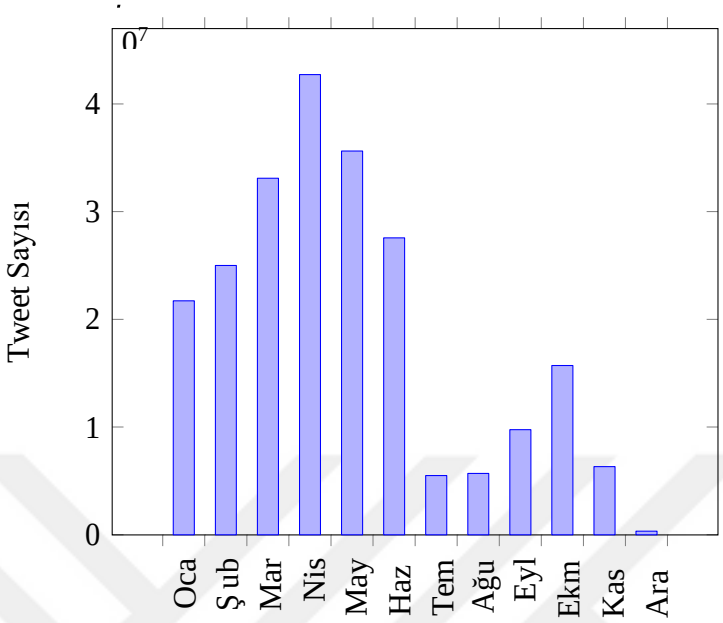
- AS-3: İddiaları teyit gerektiren olarak etiketlemenin ana gerekçeleri nelerdir?

TrClaim-19 veri kümesinin oluşturulması sırasıyla şu şekildedir: Bölüm 3.1 tweetlerin toplanmasını, Bölüm 3.2 etiketlenecek tweetlerin seçilmesini, Bölüm 3.3 etiketleme sürecini, Bölüm 3.4 veri kümesi üzerinde yapılan analizi ve Bölüm 3.5 bu veri kümesi kullanan referans modelleri anlatmaktadır.

3.1. Tweetlerin Toplanması

Yanlış bilgilerin yayılmasında sosyal medya platformları önemli bir rol oynamaktadır. Bu nedenle, Twitter üzerinden yayılan iddialara odaklanılmıştır. 1 Ocak 2019 - 31 Aralık 2019 tarihleri arasında Twitter API kullanılarak teyit gerekti- ren iddiaların olduğu tweetleri edinmek için Türkçe tweetler toplanmıştır. Önemli konulardaki tweetleri toplamak için anahtar kelime izleme yaklaşımı kullanılmıştır. 2019'da Türkiye'de yerel seçimler, İstanbul'da deprem, Suriye'de askeri operasyon ve diğerleri gibi birçok büyük olay yaşanmıştır. Bu nedenle, bu önemli olaylarla ilgili olarak "secim", "terör", "ekonomi", "multeci", "Suriye", siyasetçi isimleri (örneğin, "Erdoğan") ve diğerleri gibi anahtar kelimeler kullanılmıştır. Sabit bir anahtar kelime listesi yerine değişken bir anahtar kelime listesi kullanılarak yeni bir önemli olay olduğunda liste güncellenebilmiştir. Örneğin, 26 Eylül 2019 İstanbul depremi sonrasında anahtar kelime listesine "deprem" kelimesi eklenmiştir.

Tweet toplama sürecinin sonunda 225 milyon civarında Türkçe tweet edinilmiştir. Her ay toplanan tweetlerin dağılımı **Şekil 3.1.1**'de gösterilmektedir. Yerel seçimler, 31 Mart ve 24 Haziran'da yapıldığı (İstanbul için tekrarlanan seçimler) ve Twitter siyasi tartışmaların ana platformlarından biri olduğu için Ocak ve Haziran ayları arasında toplanan tweet sayısı diğer aylara göre çok daha fazla olduğu görülmektedir.



Şekil 3.1.1: 2019 Yılının Her Ayında Toplanan Tweet Sayısı.

3.2. Tweetlerin Seçilmesi

Toplanan çok fazla tweet arasından etiketlenebilecek sayıda tweet seçilmesi gerekmektedir. Rastgele tweetlerin seçilmesi, veri kümesinde doğruluğu teyit gerektirmeyen tweet oranının çok yüksek olmasına neden olabilir. Bilgi eris¸im alanındaki test oluşturma metodolojisinden [36] esinlenerek, ilk önce doğruluk kontrolü yapan web siteleri tarafından doğruluğu kontrol edilen iddialar toplanmıştır (Bölüm 3.2.1) ve ardından her bir iddia, kendine benzer tweetleri bulmak için bir arama sorgusu olarak kullanılmıs¸tır (Bölüm 3.2.1). İddialara benzer tweetler kullanmak, aynı zamanda aras¸tırma sorusu AS-2 için cevap aramamıza da olanak tanımaktadır.

3.2.1 İddiaların Toplanması

Koleksiyonumuzun nitelikli olması için, doğruluk kontrol metodolojisinin şeffaflığı ve tarafsızlığı gibi çeşitli konularda doğruluk kontrol web sitelerini denetle-

Algorithm 1 Tweetlerin Seçilmesi

Girdi: Tweetler T , Iddialar I
Çıktı: Seçilen Tweetler ST

```
T için indeks oluşturur (index)
ST ← []
for each iddia in C do
    ranked ← BM25(index,iddia)
    ST ← ST ∩ ranked[0]
    i ← 1
    count ← 0
    while i < ranked.size() do
        S ← LDF(ranked[i],ranked[i − 1])
        if score < threshold then
            ST ← ST ∩ ranked[i]
            count ← count +1
            if count == 3 then
                break
        i ← i +1
return ST
```

yen International Fact Checking Network (IFCN)[1] tarafından onaylanan doğruluk kontrolü yapan web siteleri dikkate alınmıştır. Bu yüzden, IFCN tarafından onaylanmış Türkiye'deki tek doğruluk kontrol siteleri olan Teyit[2] ve Doğruluk Payı[3] (DP) tarafından doğruluğu kontrol edilen iddialar toplanmıştır. Teyit, genellikle sosyal medya platformlarında yayılan iddialara odaklanırken, DP ise politikacılar tarafından ortaya atılan iddiaların doğruluğunu araştırmaktadır. Şubat 2015 ile Ocak 2020 arasında Teyit ve DP tarafından doğruluğu kontrol edilen tüm iddialar toplanmıştır. Toplanan tweetler 2019'da yayınlanmış olsa da, iddialar tarihlerine göre filtrelenmemiştir. Bunun nedeni, insanların aynı iddiayı birden çok kez öne sürebilmeleridir. Örneğin, birçok insan 1969'daki Ay'a ayak basmanın sahte olduğuna inanmaktadır ve bu iddia onlarca yıldır ortalıkta dolaşmasıdır.

[1] www.poynter.org/ifcn/
[2] www.teyit.org
[3] www.dogrulukpayi.com

Sözel olarak analiz edilebilecek iddialara odaklanıldığı için resim veya video içeren iddialar dikkate alınmamış¸tır. Toplamda, 573'ü DP'den ve 192'si Teyit'ten olmak üzere 765 iddia edinilmis¸tir.

3.2.2 Etiketlenecek Tweetlerin Seçilmesi

Etiketlenecek tweetlerin seçiminde üç ana hedef bulunmaktadır: 1) etiket dağıtımı açısından dengeli bir test koleksiyonu olus¸turmak, 2) konu ve dilbilim özellikleri açısından çes¸itli tweetlere sahip olmak ve 3) bazı tweetlerin, aras¸tırma sorusu AS-2 için uzmanlar tarafından doğruluğu kontrol edilen iddialar içerdiğinden emin olmak.

Tweet seçiminde kullanılan algoritma, Algoritma 1'de sunulmus¸tur. Öncelikle, Lucene arama motoru kütüphanesi [Satır 1] kullanılarak tüm tweetler kelimeleri köklerine ayrılmadan indekslenmis¸tir. Bu is¸lem sayesinde indekslemenin yanı sıra retweetler nedeniyle tweet koleksiyonlarında çok fazla sayıda bulunan aynı tweet de ortadan kaldırılmıs¸tır.

Her iddia için iddianın kendisini bir arama sorgusu olarak kullanılmıs¸tır ve edinilen tweetler, BM25 sıralama algoritması ile benzerlik puanlarına göre sıralanmıs¸tır [Satır 4]. Bununla birlikte, en üst sıralarda yer alan tweetleri almak yerine, edinilen iddialarla ilgili ancak birbirinden farklı tweetleri bulmak için iki as¸amalı bir süreç uygulanmıştır. Bunun nedeni, birçok tweetin tam olarak aynı olmadığının ancak çok benzer olduğunun gözlemlenmesidir. Örneğin, iki tweet aynı mesaja sahip olabilir, ancak bunlardan birinin ek bir URL'si, hashtag'i ve/veya emojisi bulunabilmesidir.

İlk olarak etiketlenecek en üst sıradaki tweet seçilmis¸tir [Satır 5]. Ardından, en üst sıradaki tweet'lerden bas¸layarak, arka arkaya sıralanmıs¸ tweetlerin metinsel benzerlikleri hesaplanmaktadır [Satır 9]. Tweetlerin benzerliği önceden tanımlanmış bir eşik değerinden düşükse, ilgili tweet, daha düşük bir sıraya sahip tweet, seçilen tweetler listesine eklenmektedir.

İki tweet arasındaki metinsel benzerliği hesaplamak için, önce kullanıcı adları, URL'ler ve alfanümerik olmayan karakterler kaldırılır ve ardından Levenshtein Mesafesine göre yaklas¸ık benzerlik hesaplanmaktadır. Deneysel olarak, benzerlik puanı eşiği 0,80 olarak belirlenmiştir. Bu yöntem kullanılarak her bir iddia için üç tweet seçilmis¸tir [12-14. Satırlar].

Yukarıda açıklanan algoritmayı kullanarak DP'den edinilen iddialar için 1714, Teyit'ten edinilen iddialar için 573 tweet üzere toplamda 2287 tweet elde edilmis¸tir[4].

3.3. Etiketleme Süreci

Etiketleme sürecine tez yazarı ve danıs¸manı ile 7 gönüllü dahil olmus¸tur. Etiketleyicilerin tümünün anadili Türkçedir ve yaş aralığı 22 ile 35 arasındadır. Etiketleyiciler, bilgisayar bilimi, diş hekimliği, tarih, hukuk ve işletme dahil olmak üzere çes¸itli disiplinlerde derecelere sahiptir. Farklı geçmis¸lere sahip etiketleyicilerin olması, doğruluğu teyit gerektiren iddialarla ilgili farklı düşünceler ve bakış açıları edinilmesine olanak tanımaktadır.

Etiketleme sürecine bas¸lamadan önce, her bir gönüllüye veri kümesi ve etiketleme görevi açıklanmıs¸tır. Etiketleme arayüzünde, katılımcılar, etiketlenecek tweetleri ve bunları edinmek için kullanılan ilgili iddiaları görmektedir. Etiketleme görevinde aşağıdaki üç soru sorulmaktadır.

- Tweet, iddiayla alakalı mı?
- Tweetin do ğruluğu teyit gerektiren bir iddia içerdiğini düşünüyor musunuz?
- Tweetin teyit gerektiren bir iddiası oldu ğunu düşünüyorsanız gerekçeniz nedir?

[4]arama uygulaması, bazı iddialar için üçten daha az tweet dönmüs¸tür

Çizelge 3.3.1: TrClaim-19 Hakkında Genel ˙Istatistikler

Toplanan Tweet Sayısı	225M
Etiketlenen Tweet Sayısı	2287
Teyit gerektiren ˙Iddia Sayısı	875
Gerekçe Kategorilerinin Sayısı	26

Etiketleyiciler, teyit gerekliliğine ilişkin yargılarının arkasına gerekçelerini yazmak için serbest bir metin formu kullanmıs¸lardır. Yalnız, daha sonrasında veri kümesindeki gerekçeleri daha iyi analiz etmek için etiketleyicilerden gerekçe tanımlarında tutarlı olmaları için ellerinden geleni yapmalarını ve aynı gerekçe için aynı metin ifadesini kullanmalarını talep edilmis¸tir.

Her bir tweet, üç ayrı etiketleyici tarafından etiketlenmis¸tir. Katılımcılar, 3 gruba ayrılmıs¸tır. Sonrasında, her bir etiketleme grubu ayrı bir tweet kümesini etiketlemis¸tir. Tweet kümelerinin boyutları, bu etiketleme görevini gönüllü olarak yaptıkları için her bir etiketleme grubunun müsaitlik durumuna göre belirlenmis¸tir. Gruplara atanan tweet sayısı 384, 427 ve 1476'dır. Etiketleme sürecinde önyargıyı önlemek için, katılımcılar etiketleme is¸lemi bitene kadar bas¸kalarının etiketlerini görmemis¸lerdir. Toplamda 2287x3 = 6861 ilgililik kararı ve teyit gereklilik kararı toplanmıs¸tır.

Çizelge 3.3.2: Etiket Dağılımı. İlgililik Kararları Çoğunluğun Kararına Göre Belirlenmis¸tir. Teyit gereklilik (TG) oranı, her bir tweet için yapılan etiketlemedeki "teyit gerektiren" etiketi oranını belirtmektedir.

˙Ilgililik	**TG Oranı**	**Tweet Sayısı**		
		Teyit	**Doğruluk Payı**	**Toplam**
˙Ilgili	0/3	32	46	78
	1/3	85	182	267
	2/3	141	246	387
	3/3	142	100	242
İlgili Değil	0/3	84	594	678
	1/3	52	337	389
	2/3	29	182	211
	3/3	8	27	35

Çizelge 3.3.3: Uzman Tarafından Doğrulaması Yapılan İddilarla İlgili Farklı Teyit Gereklilik Oranlarına Sahip Örnek Tweetler

Tweet	**TM Oranı**
Yüksek teknolojili ürünlerin imalattaki payı yüzde 3'e geriledi	0/3
Bugüne kadar Suriyeli sığınmacılar için harcanan para 40 milyar Dolar	1/3
Trafik kazalarında son on yılda 52 bin 95 kis¸i yas¸amını yitirdi	2/3
Avrupa'da genç işsizliği yükselen iki ülkeden biri Türkiye	3/3

3.4. Analiz

Çizelge 3.3.1, TrClaim-19 ile ilgili genel istatistikleri göstermektedir. Çoğunluk oylamasına dayalı olarak her bir tweetin ilgililik ve teyit gereklilik yargıları bir araya getirildiğinde, tweetlerin 974'ü (= 78 + 267 + 387 + 242) ilgili (koleksiyonun % 42'si) ve 875'i (= 387 + 242 + 211 + 35) teyit gerektiren (koleksiyonun % 38'i) bulunmus¸tur.

AS-1: Uzman olmayanlar arasında iddiaların teyit gerekliliği konusunda anlas¸ma düzeyi nedir?

Topladığımız etiketler, uzman olmayan bilgi denetçilerinin, çalışmamızdaki etiketleyicilerin, iddiaların teyit gerekliliği konusunda büyük ölçüde aynı fikirde olmadığını göstermektedir. Çizelge 3.3.2'de, etiketleyicilerin 1033 (= 78 + 242 + 678 + 35) tweetin (tüm koleksiyonun % 45'i) teyit gerekliliği konusunda tamamen hemfikir olduğunu, ancak kalan 1254 tweet için (tüm koleksiyonun % 55'i) birbirine tamamen katılmadığı gözlemlenmektedir. Teyit gereklilik değerlendirmeleri için Fleiss kappa puanı 0,23'tür ve Fleiss vd.[12] göre bu durum "adil uzlaşma" anlamına gelmektedir. Referans noktası olarak, ilgililik değerlendirmeleri için Fleiss'in kappa puanı 0,61 ile "belirgin bir uzlas¸ma" anlamına gelmektedir. Bu, iddiaların teyit gerekliliğine karar vermenin, ilgililiğine karar vermekten daha öznel bir görev olduğunu göstermektedir.

Çizelge 3.4.1: TrClaim-19 İçinde En Yaygın Belirtilen Gerekçe Grupları ve Her Grup İçin Örnek Cümle. Her grupun veri kümesinde kaç kere belirtildiği parantez içinde verilmiştir.

Gerekçe Grubu	**Etiketleyiciler Tarafından Verilen Gerekçeler**	**Örnek Tweet**
Ekonomi (508)	"ekonomi", "ekonomik durumun tespiti", "hükümet kararlarının ekonomiye etkisi"	Türkiye'de her 10 kis¸iden 7'si borçlu, en yoksul kesimin toplam gelirden aldığı pay ise sadece yüzde 6,1
Politika (376)	"seçim güvenliği", "seçim sonrası toplumsal etkileri", "seçim süreci ve etkileri", "siyaset ve topluma etkileri", "siyasi", "Kurumlara/kis¸ilere dair algıyı kötü etkileyebilir", "siyasi propaganda", "İnsanların politik görüs¸ünü etkileyebilir", "partiye yönelik", "belediye hizmetleri"	CHP'nin Kılıçdaroğlu saldırısıyla ilgili verdiği Meclis araştırma önergesi reddedildi...
Toplumsal (287)	"toplumsal sorunlar", "toplumsal cinsiyet eşitsizliği", "Toplumun önem verdiği bir konu ile alakalı", "insanların psikososyal durumunun tespiti", "sosyal", "toplumsal ulusal"	Kadın is¸ gücüne katılım oranını yüzde 34,1'e çıkardık
Genel (187)	"gerçekles¸mis¸ durumun tespiti"	82 milyon nüfus ile dünyanın en kalabalık 18. ülkesi olan Türkiye aynı zamanda dünyanın 19. büyük ekonomisi.
Uluslararası (58)	"uluslararası politika", "dıs¸ gündem", "evrensel", "türkiyedeki olayların farklı olaylara etkisi", "uluslararası kıyaslama"	ABD Dıs¸is¸leri Bakanı Mike Pompeo: "Trump gerekirse Türkiye'ye kars¸ı askeri güç kullanmaya hazır."
İlginç (41)	"iddia ilginç", "tuhaf"	Avrupa Merkez Bankası üzerinde Atatürk portresi olan banknotlar bastırdı.
Güvenlik (35)	"emniyet", "askeri", "terör"	Daha dün milli tankını, tüfeğini, helikopterini, denizaltısını, İHA ve SİHA'sını üreten Türkiye hayaldi, Erdoğan ile gerçek oldu? Yerli savunma sanayi (silah) yüzde 15'lerden yüzde 75'lere çıktı.
Şüpheli İddialar (33)	"s¸üpheli", "eksik bilginin tamamlanması, kontrolü"	Cumhurbaşkanı Erdoğan, 2019'un ilk 5 ayında örtülü ödenekten tam 1 milyar 6 milyon 621 bin lira harcamıs¸.
Eğitim (33)	"Eğitim"	Üniversite mezunu is¸siz 1 milyonu as¸tı
Sağlık (28)	"sağlık", "kaza", "iş güvenliği"	Dikkat arttı, is¸ kazalarında azalma var

Daha önce de bahsedildiği gibi, etiketleyiciler üç gruba ayrılmıştır. Gruplar arasında etiketleme anlaşması düzeylerinde herhangi bir fark olup olmadığını görmek için, her grup için ayrı ayrı Fleiss kappa puanı da hesaplanmıs¸tır. Fleiss kappa puanları, her grubun teyit gereklilik yargıları için 0.17 ile 0.32 arasındadır ve bu da gruplar arasında kayda değer bir fark olmadığını göstermektedir.

Sonuçlar, aras¸tırma sorusuyla (AS-5) ilgili olarak uzman olmayan kis¸ilerin iddiaların teyit gerekliliği konusunda fikir ayrılığına düştüğünü göstermektedir. Bu nedenle, mevcut veri kümelerinde olduğu gibi ikili etiketler kullanmak yerine, derecelendirilmiş etiketlerin bu görev için daha uygun olabileceği çıkarımı yapılabilir. TrClaim-19'daki etiketlerin, her tweet için teyit gereklilik oranları kullanılarak kolayca derecelendirilmiş etiketlere dönüştürülebileceğini belirtilmektedir.

AS-2: Uzman olmayanların iddiaların teyit gerekliliği konusunda uzmanlardan farklı fikirleri var mı?

TrClaim-19'daki ilgili tweetler, bu aras¸tırma sorusuna cevap aramak için faydalı olabilir. Teyit ve DP'den topladığımız iddiaların uzmanlar tarafından doğruluğu kontrol edilmis¸tir. Bu nedenle, uzmanların bu iddiaları teyit gerektiren buldukları rahatlıkla varsayılabilir. Çizelge 3.3.2'de, 78 ilgili tweetin teyit gereklilik oranının 0/3 olduğu görülmektedir, yani tüm etiketleyiciler uzmanlarla aynı fikirde değildir. Dahası, 267 tweet için etiketleyicilerin üçte ikisi, tweetleri uzmanların aksine teyit gerektiren bulmamıs¸tır. Genel olarak, % 36 (= 78 + 267 + 387 + 242 78 + 267) oranında, etiketleyicilerin uzman doğruluk denetleyicileriyle aynı fikirde olmadığı gözlemlenmis¸tir.

Uzmanlar ve uzman olmayanlar arasındaki bu anlaşmazlığa daha iyi anlamak için, Çizelge 3.3.3, farklı teyit gereklilik oranlarına sahip ilgili örnek tweetlerini göstermektedir. Tüm etiketleyicilerin karşılığındaki iddiayla alakalı olduğuna karar verdiği tweetler seçilmiştir. Çizelgede, etiketleyicilerin hiçbirinin yüksek teknolojili ürünlerin üretimdeki payına ilişkin iddiayı dikkate almadığını görülmektedir. Bunun nedeni, endüstri hakkındaki bu iddianın yaşamlarını doğrudan etkilememesi olabilir. Öte yandan, tüm katılımcılar, genç işsizliği iddiasını teyit gerekliliği olarak değerlendirmiştir. Bunun nedeni, etiketleme yapanların genç olması ve bu nedenle genç işsizliğiyle ilgilenmesi olabilir.

Etiketleme sürecimiz sırasında etiketleyiciler, ilgili iddiaların uzmanlar tarafından doğruluğunun kontrol edildiğini bilmektedir. Bu, potansiyel olarak kararlarını etkileyebilmektedir. Bu potansiyel önyargının analizini gelecek çalıs¸ma olarak incelenebilir. Bununla birlikte, AS-2 ile ilgili olarak, uzmanların bu iddiaları kontrol ettiğini bilmelerine rağmen, etiketleyicilerin çoğu durumda uzmanlarla aynı fikirde olmaması dikkat çekicidir.

AS-3: Bir iddiayı teyit gerektiren olarak etiketlemenin ana nedenleri nelerdir?

Etiketleyiciler, tweetleri 2683 kez teyit gerektiren olarak değerlendirmiştir (çoğunluğun kararı uygulanmadan) ve bu kararlar için gerekçeler sağlanmıştır. Gerekçeleri nasıl tanımlamaları gerektiği hususunda herhangi bir sınırlama koyulmadığı için toplamda 71 farklı metin gerekçe olarak belirtilmiştir. Etiketleyicilerin aynı veya benzer gerekçeler için farklı metinler kullandığını gözlemlenmiştir (örneğin, "mülteci hakları", "mülteciler" ve "mülteci sorunları"). Bu nedenle, tüm gerekçeler elle incelenerek 26 kategori altında gruplandırılmıs¸tır.

Bazı gerekçelerin sadece birkaç kelime olduğunun (örneğin, "insan hakları", "sağlık") gözlemlenmiş olması iddianın önemli bir konuyla ilgili olduğu için teyit gerekliliği olduğunu öne sürmektedir. Öte yandan bazı etiketleyiciler, "İddia ilginç", "Bu iddia nedeniyle insanlar hayatları hakkında önemli bir karar verebilir", "kis¸i veya kurumlara yönelik algıyı olumsuz etkileyebilir" gibi açık gerekçe ifadeleri sunmus¸lardır.

Çizelge 3.4.1, en sık belirtilen 10 gerekçe grubunu, etiketleyiciler tarafından sağlanan gerçek gerekçe metinlerini ve her grup için örnek bir tweeti göstermektedir. Geriye kalan gerekçe grupları s¸u s¸ekildedir: ekoloji (23), mülteciler (21), tarım (17), ulusal değerler (16), skandallar (13), tarihi olaylar (12), inkarlar (12), insan hakları (11), bilim (10), turizm (8), medya (7), altyapı (7), siyasi olmayan konularda etkisi(6), s¸iddet (2) ve kültür (1).

Gerekçeler, teyit gerektiren iddialar için yararlı bilgiler sağlamaktadır. Başlıca gözlemler, aşağıdaki gibidir.

˙Ilk olarak, bir iddianın konusu, onu teyit gerektiren kılmak için önemli bir faktördür. Özellikle, iktisat, siyaset ve sosyal konular, teyit gerektiren iddialarda en yaygın konulardır.

İkinci olarak, insanların bir iddiayı teyit gerektiren olarak değerlendirmesi için farklı gerekçeleri olabilmektedir. Örneğin, üç etiketleyici aşağıdaki tweeti teyit gerekliliği olarak değerlendirmiş ancak farklı gruplarda (sağlık, politika ve sos- yal konular) gerekçeler sağlamışlardır: "Pazartesi günü tıbbi muayene ve ilaç için katkı payları sırasıyla % 60 ve % 70 artacak". İddia açıkça sağlıkla ilgili olsa da, diğer etiketleyiciler onu siyaset ve sosyal yaşam üzerindeki etkisi açısından teyit gerektiren bulmaktadır.

Üçüncüsü, etiketleyiciler, gerekçelerinde sık sık olumsuz sonuçları veya sorunları ifade etmişlerdir (örneğin, "ekolojik sorunlar", "önemli bir konudaki kuralların ihlali", "yasadıs¸ı eylem", "mültecilerin sorunları" ve "adaletsizlik"). Bu, olumsuz konularla ilgili iddiaların, olumlu konularla ilgili iddialardan daha fazla teyit gerektiren olduğunu göstermektedir.

Dördüncüsü, toplanan birçok gerekçe, "insanların siyasi durus¸unu etkileyebilir", "tarihte etkili bir kişiye yönelik algıyı değiştirebilir", "insanlar bu iddiaya dayanarak önemli bir karar verebilir" gibi iddiaların olası etkileri hakkındadır. Bu nedenle, teyit gerektiren iddiaları önceliklendirirken iddiaların etkisi de dikkate alınması gerekmektedir.

Son olarak, "ulusal değerler" ve "toplum için önemli bir konu hakkında" gibi gerekçeler, bazı iddiaların teyit gerekliliğinin evrensel olmadığını, ancak belirli bir ulus / ülke için kontrol değerinde olduğunu göstermektedir. Bu nedenle, teyit gerektiren iddia önceliklendirme modellerinde, her ülkenin ulusal sorunları da dikkate almalıdır.

3.5. Referans Sonuçlar

Bu bölümde, gelecekteki çalıs¸malara yardımcı olmak amacıyla dört farklı referans modelin performans sonuçlarını sunulmus¸tur. Teyit ve DP'den toplanan 765 iddia içinden 635 iddia seçilmiş ve bu iddialarlar ilgili 1900 tweet eğitim için kullanılmıs¸tır. Kalan 387 tweet ise test için kullanılmıs¸tır. Kullanılan referans modeller aşağıdaki gibidir.

- **MBERT:** CTL20-ED'de en iyi performans gösteren modeller, BERT modelinin [9] türevlerinin kullanmıştır [6]. Eğitim verilerini kullanarak BERT'in çok dilli versiyonuna (MBERT) hassas ayar yapılmıs¸tır.
- **BERTurk:** Pires vd. [32], MBERT'in performansının cümle içinde farklı kelime sıralarına sahip diller için düşebileceğini belirtmişlerdir. Türkçe de ˙Ingilizce'den farklı kelime sıralamasına sahiptir. Bu nedenle, sadece Türkçe metinler kullanılarak önceden eğitilmiş tek dilli bir BERT modeli olan BERTurk'e [33] hassas ayar yapılmıs¸tır.
- **LR-BOW:** Kelime torbasını öznitelik olarak kullanan bir lojistik regresyon modeli eğitilmiştir. Öncesinde bazı ön işleme tekniklerini uygulanmıştır: küçük harfe çevirme, alfabetik olmayan karakterleri kaldırma, NLTK[5] ile sık kullanılan kelimeleri çıkarma ve kelimeleri köklerine ayırma[6]. Yine NLTK kullanarak tweetler kelimelere ayrılmıs¸ ve öznitelik olarak kullanılacak 6157 kelimeden olus¸an kelime torbası olus¸turulmus¸tur.
- **SVM-BOW:** Aynı s¸ekilde kelime torbası öznitelik olarak kullanılarak SVM modeli eğitilmiştir.

LR-BOW ve SVM-BOW modelleri, Scikit[7] kütüphanesi kullanılarak varsayılan parametrelerle eğitilmiştir. Tüm modeller, ikili ve çoklu sınıflandırma olarak iki şekilde eğitilmiştir. İkili sınıflandırmada, tweetlerin teyit gerekliliği çoğunluğun

[5] www.nltk.org/
[6] https://snowballstem.org
[7] https://scikit-learn.org

Çizelge 3.5.1: Referans Modellerin TrClaim-19 Veri Kümesinde Değerlendirme Sonuçları. En iyi sonuçlar **koyu** gösterilmis¸tir.

Model	AP	P@1	P@5	P@10	P@30	R-P	nDCG
M-BERT	.5508	0.0000	**.6000**	**.6000**	.6667	.5548	.8643
BERTurk	**.5810**	**1.0000**	**.6000**	**.6000**	**.7000**	**.5685**	.8756
LR-BOW	.3609	**1.0000**	.2000	.2000	.2667	.3245	**.8815**
SVM-BOW	.3716	**1.0000**	.2000	.2000	.3333	.3444	.8372

kararı uygulanarak "teyit gerektiren" ve "teyit gerektirmeyen" etiketleri kullanılmıs¸tır. Çoklu sınıflandırmada ise, her tweet için toplam teyit gerektiren etiket sayısını belirten dörtlü bir etiket ölçeği (0, 1, 2 ve 3) kullanılmıştır. Referans modeller kullanılarak iddialar teyit gerekliliklerine göre önceliklendirilmis¸tir. İkilisınıflandırma sonuçları, AP, RP ve P@k ölçütleriyle değerlendirilirken, çoklu sınıflandırma sonuçlarının değerlendirilmesinde nDCG puanları kullanılmıştır.

Çizelge 3.5.1, referans modeller için değerlendirme sonuçlarını göstermektedir. İkili sınıflandırmada tüm değerlendirme ölçütlerine göre, en başarılı model BERTurk'tür. Beklendiği gibi, BERT tabanlı modeller çoğu durumda LR-BOW ve SVM-BOW modellerinden daha başarılıdır. Çoklu sınıflandırmanın değerlendirildiği nDCG ölçütüne göre, tüm modellerin sonuçları genel olarak birbirine yakın olsa da en bas¸arılı model LR-BOW'dur.

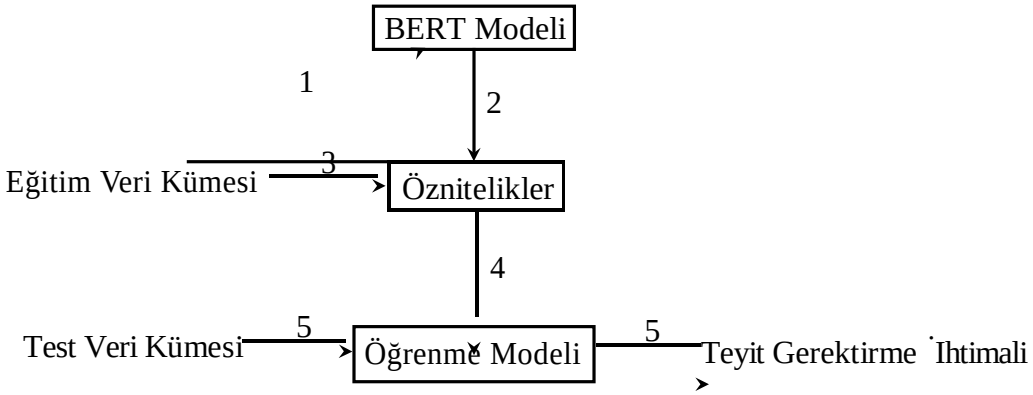
4 TEY İTGEREKTİREN İDDİALARIN ÖNCELİKLENDİRİLMESİ

Tezin bu bölümünde, AS-4: İddialar teyit gerekliliklerine göre ey iyi nasıl önceliklendirilebilir? sorusuna cevap aranmaktadır. Bunun için, BERT modeline hassas ayar yapılmıs¸ ve bu modelin tahmin sonuçları ve farklı öznitelikler kullanılarak bir lojistik regresyon modeli önerilmis¸tir. Kullanılan öznitelikler arasında kelime vektörleri, karşılaştırma ifadelerinin varlığı, yerel bölgeye özgü tartışmalı konular ve diğerleri bulunmaktadır. Modelimiz, CTL'18 ve CTL'19 veri kümelerinde sırasıyla 0,255 ve 0,176 MAP puanlarına ulas¸arak, ilgili veri kümelerini kulla- nan CTL katılımcıları, Claim-Buster [21], BERT, XLNET [40] ve Lespagnol vd. [25] 'in modelleri de dahil olmak üzere tüm en bas¸arılı modellerden daha iyi performans göstermis¸tir.

4.1. Önerilen Yöntem

Teyit gerektiren iddiaların önceliklendirilmesi için yapılan çalıs¸maların bir kısmı, etkili öznitelikleri araştırırken, diğerleri de dönüştürücü (transformer) modellerin en iyi nasıl hassas ayar yapılacağına odaklanmaktadır. Bizim yaklaşımımız ise bu iki farklı aras¸tırma yönünü birles¸tirmeyi amaçlamaktadır.

Önerilen yöntemin aşamaları, Şekil 4.1.1'de gösterilmiştir. Önerilen yöntem, hassas ayarlı bir dönüştürücü modelinin tahminini aşağıda açıklanan diğer özniteliklerle birlikte kullanan karma bir yaklaşımdır. CTL tarafından sağlanan veri kümeleri tartıs¸ma dokümanlarını içermektedir. Tartıs¸maların interaktif söylem yapısı nedeniyle, birçok kesilmis¸ veya eksik cümle bulunmaktadır. Bu nedenle ilk önce, bir veya iki kelimeli ifadeler filtrelenmis¸tir. Ardından ifadeleri sıralamak için de-



Şekil 4.1.1: İddiaların Teyit Gerektirmelerine Göre Önceliklendirilmesi İçin Önerilen Yöntem. 1) Öncelikle, BERT modeli eğitim veri kümesi kullanılarak hassas ayarlanır ve 2) tahmin sonuçlarından öznitelik olus¸turulur. 3) Diğer öznitelikler çıkartılır. 4) Öznitelikler kullanılarak öğrenme modeli eğitilir. 5) Eğitilen öğrenme modelinden test veri kümesi için teyit gerektirme ihtimalleri elde edilir.

netimli modelimiz uygulanmıs¸tır. LR, SVM, RF, MART [14] ve LambdaMART [39] dahil olmak üzere çeşitli öğrenme modelleri incelenmiştir. Kullanılan öznitelikler s¸u s¸ekildedir.

- **MBERT:** İlgili eğitim veri kümeleri kullanılarak MBERT modeline hassas ayar yapılmıştır. Ardından, bu modelin tahmin değeri özniteliklerden biri olarak kullanılmıs¸tır.
- **Kelime Vektörleri (KV):** Anlamsal ve sözdizimsel olarak benzer olan kelimeler vektör alanında yakın olma eğilimindedir ve iddialar arasındaki benzerliklerin bulunmasına yardımcı olmaktadırlar. Her bir cümle, vektörü mevcut olan kelimelerinin vektörlerinin ortalaması olarak temsil edilmektedir. Kelime vektörleri, 300 boyutlu vektörlere sahip önceden eğitilmiş word2vec [28] modelinden edinilmis¸tir.
- **Tartıs¸malı Konular (TK):** Tartıs¸malı konular hakkındaki cümleler, teyit gerektiren iddiaları içerebilmektedir. Lespagnol vd. [25], Wikipedia makalesi olan "Wikipedia: Tartıs¸malı konuların listesi" nden derlenen tartıs¸malı konuların bir listesini çalıs¸malarında kullanmıs¸lardır. Ancak kullandıkları

liste, "Lübnan", "Çernobil" ve "İspanya İç Savaşı" gibi mevcut ABD medyasında çok sınırlı yer alan birçok tartışmalı konuyu kapsarken, mevcut veri kümeleri güncel ABD siyaseti hakkındadır.

Bir konudaki tartışmanın topluma bağlı olduğunu düşünülmektedir. Örneğin, ABD'li politikacılar göçmenler için farklı politikalar önermektedirler ve bu politikalar, göçmenler ile destekçileri arasında tartışmalara yol açmaktadır. Öte yandan, ABD iç siyaseti, Akdeniz'deki mülteci kriziyle Avrupa ülkelerine göre çok daha az ilgilenmektedir. Bu nedenle, Meksikalı göçmenlerle ilgili bir iddia, ABD'de yaşayan insanlar için doğruluğu teyit gerektiren olabilirken, mültecilerin Avrupa'ya ulaşmak için tehlikeli bir yol izlediklerine dair iddiaları önemsiz bulabilmektedirler. Buna karşılık, Avrupa'da yaşayan insanlar ikinci vakayı teyit gerektiren, ilk vakayı ise teyit gerektirmeyen olarak değerlendirebilir.

Bir konudaki tartışmalar zamanla değişebilmektedir. Örneğin, Wikipedia listesinde de olan "Soğuk Savaş", 1991'de Sovyetler Birliği'nin dağılmasından önce ABD siyasetinde en çok tartışılan konulardan biridir. Ancak günümüzde ABD medyasında nadiren yer almaktadır. Bu nedenle, dünya çapında ve tarihte tartışmalı herhangi bir konu yerine, kullanılan verilerle ilgili tartışmalı konuların kullanması öngörülmektedir.

İlk olarak, mevcut ABD siyasetinden göç, silah politikası, ırkçılık, eğitim, İslam, iklim değişikliği, sağlık politikası, kürtaj, LGBT, terör ve Afganistan ve Irak'taki savaşlar dahil 11 ana konu belirlenmiştir. Her konu için, ilgili kelimeleri belirlenmiş ve kelime vektörlerini kullanarak bu kelimelerin vektör ortalamaları hesaplanmıştır. Örneğin, göçmenlik konusunda "göçmenler", "yasadışı", "sınırlar", "Meksikalı", "Latin" ve "Hispanik" kelimeleri kullanılmıştır.

Bu 11 boyutlu öznitelik grubunda, kelime vektörleri kullanılarak cümleler ve her konu arasındaki kosinüs benzerliğini hesaplanmıştır. NLTK ile sık kullanılan kelimeleri (stop words) hariç tutularak cümleler için kelime vektörlerinin ortalamaları kullanılmıştır.

- **Kars¸ılas¸tırma I˙fadeleri (KI˙):** Politikacılar sık sık kendilerini bas¸kalarıyla kars¸ılas¸tıran cümleler kullanırlar. Çünkü, halkı rakiplerinden daha iyi olduklarına ikna etmeye çalıs¸ırlar. Bu nedenle, siyasi konus¸malardaki kars¸ılaştırmalar, insanların oy verme kararlarını etkileyebilir ve bu yüzden, doğruluğunu kontrol etmek önemli olabilmektedir. Bu nedenle, bu öznitelik için, cümlelerde kars¸ılas¸tırma sıfatlarının ve zarflarının sayısı kullanılmaktadır.

- **Özel Kelime Listesi (KL):** Belirli kelimeler,teyit gereklilik hakkında önemli bilgiler verebilmektedir, çünkü 1) önemli bir konuyla ilgili olabilir (örneğin, "işsizlik"), 2) sayısal bir değeri temsil edebilir ve cümlenin gerçekliğini artırabilir (örneğin, "yüzde") ve 3) semantik, iki durum arasındaki bir karşılaştırmayı belirtebilirler (örneğin, "artış" ve "azalma"). Bu nedenle, ilk olarak CTL'18 ve CTL'19 eğitim veri kümeleri analiz edilerek 66 kelime belirlenmis¸tir. Bu öznitelikte, seçilen kelimelerin bas¸sözcükleri (lemma) ile ilgili cümledeki kelimelerin başsözcükleri arasında bir örtüşme olup olmadığını kontrol edilmektedir.

- **Yüklem Zaman Kipleri (YZ):** Gelecekle ilgili iddiaların teyit edilemez, ancak s¸imdiki zaman veya geçmis¸le ilgili iddiaları teyit edilebilir. Bu nedenle, cümlelerin yüklem zamanı, iddiaların teyit gerekliliği için etkili bir gösterge olabilir. Bu öznitelik vektörü, cümlelerin yüklemindeki çes¸itli zaman çekimlerinin varlığına göre belirlenmiştir.

- **POS Etiketleri (POS):** Bir cümle herhangi bir bilgilendirici kelime içermiyorsa, o zaman teyit gerektirme olasılığı daha düşüktür. Bir iddianın bilgi yükünü belirtmek için, isim, fiil, zarf ve sıfatların sayıları öznitelik olarak kullanılmıs¸tır.

4.2. Deneyler

Bu bölümde, deney düzeneği (Bölüm 4.2.1), sonuçları (Bölüm 4.2.2) ve geliştirilen modelin çıktıları üzerine yapılan nitel analiz (Bölüm 4.3) anlatılmaktadır.

4.2.1 Deney Düzeneği

Gerçekleştirme ayrıntıları, deneylerde kullanılan veri kümeleri, geliştirilen modelin karşılaştırıldığı referans modeller ve değerlendirme ölçütleri bu bölümde açıklanmaktadır.

4.2.1.1 Gerçekleştirilme

BERT modeline, veri kümeleri kullanılarak 2e-5 ve bir döngü ile hassas ayar yapmak için ktrain[1] kütüphanesi kullanılmıştır. Varsayılan olarak, aşırı öğrenmeyi önlemek için bir bırakma (dropout) katmanı kullanılmıştır. Tüm sözdizimsel ve anlamsal ifadelerin çıkarımları için SpaCy[2] kullanılmıştır. SVM, RF ve LR uygulamaları için Scikit kütüphanesi kullanılmıştır. Öğrenme algoritmalarının parametre ayarları aşağıdaki gibidir. SVM için varsayılan parametreler kullanılmaktadır. RF için ağaç sayısı 50, maksimum derinliği 5 olarak ayarlanmıştır. LR için çok terimli ve lbfgs ayarları kullanılmıştır. MART ve LambdaMART modelleri için RankLib[3] kütüphanesi kullanılmıştır ve ağaç ve yaprak sayısını sırasıyla 50 ve 2 olarak belirlenmiştir. İddiaları önceliklendirmek için öğrenme modellerinin aktivasyon fonksiyonu çıktıları kullanılmıştır.

[1] https://pypi.org/project/ktrain/
[2] https://spacy.io/
[3] https://sourceforge.net/p/lemur/wiki/RankLib/

Çizelge 4.2.1: Kullanılan Veri Kümeleri Detayları. Teyit gerektiren iddia oranları parantez içinde verilmis¸tir.

		CTL18	CTL19	CTL20-ED	CTL20-ET	CTL20-AR	TrClaim-19
	Dil	˙Ingilizce	˙Ingilizce	˙Ingilizce	˙Ingilizce	Arapça	Türkçe
Eğitim	**Doküman**	3	19	50	1	1	1
	Cümle	4,064	16,421	42,776	822	1500	1900
	TG	90	433	487	290	458	729
	I˙ddia	(%2.2)	(%2.6)	(%1.1)	(%35.3)	(%30.5)	(%38.4)
Test	**Doküman**	7	7	20	1	1	1
	Cümle	4,882	7,079	21,514	140	6000	387
	TG	192	110	136	60	1604	146
	I˙ddia	(%3.9)	(%1.6)	(%0.6)	(%42.9)	(%26.7)	(%37.7)

4.2.1.2 Veri Kümeleri

Bu tez çalışmasında, CTL'18, CTL'19 ve CTL'20 tarafından sağlanan İngilizce ve Arapça veri kümeleri ve Türkçe veri kümesi TrClaim-19 dahil olmak üzere altı farklı veri kümesi kullanılmıs¸tır. Bunlarla ilgili ayrıntılar Çizelge 4.2.1'de verilmis¸tir. CTL'18 ve CTL'20-ED, tartıs¸ma programlarının ve konus¸maların dökümlerinden olus¸urken, CTL'19 ayrıca basın konferansları ve gönderileri içermektedir. CTL'20-AT, "Arap Dünyasında Covid-19" ve "Sudan ve normalles¸me" gibi çes¸itli konularda Arapça tweet'lerden olus¸urken, CTL'20-ET, Covid-19 hakkında ˙Ingilizce tweet'lerden olus¸maktadır. TrClaim-19, ˙Istanbul'da yas¸anan deprem ve belediye seçimleri gibi Türkiye'de yas¸anan önemli olayları takip eden 2019'da toplanan Türkçe tweetleri kapsamaktadır. Tek öznitelikli deneylerde ve CTL katılımcıları ile kars¸ılas¸tırmalarda, adil bir kars¸ılas¸tırma yapmak için CTL ile aynı kurulum kullanılmaktadır. Ancak farklı eğitim kümelerinin etkisini araştırmak için yapılan bir sonraki bölümdeki (Bölüm 5.1) deneylerde eğitim kümesini değiştirilip test kümesi aynı tutulmaktadır. Örneğin, CTL'18 ile ilgili deneyler yapılırken, CTL'18 eğitim kümesi ve diğer veri kümeleri kullanılarak BERT modeline hassas ayar yapılmıs¸ ve ardından CTL'18 test kümesinde sonuçlar alınmıs¸tır.

Referans Yöntemler

Geliştirilen modelimiz aşağıdaki modellerle karşılaştırılmıştır.

- **Lespagnol vd. [25]:** Lespagnol vd., CTL'18'de s¸imdiye kadarki en iyi sonuçları elde etmis¸lerdir. Bu nedenle, onların çalıs¸masını referanslardan biri olarak kullanılmaktadır. CTL'19 sonuçlarını almak için, kendi kodlarını almak üzere yazarlarla iletis¸ime geçilmis¸tir. Yazarlardan "bilgi beslenme" özniteliklerinin değerleri ve kelime vektörlerinin nasıl oluşturulacağına dair talimatlar edinilmiştir. Paylaştıkları değerler kullanılarak ve verdikleri talimatlar takip edilerek deneyler yapılmıs¸tır. CTL20-ED için ise, "bilgi beslenme" özniteliklerini olus¸turmak için kendi uygulamamız kullanılmıs¸tır.
- **ClaimBuster:** CTL'18 ve CTL'19'da mevcut olmayan tartıs¸ma programları dökümlerini kapsayan bir veri kümesi üzerinde önceden eğitilmiş olan ClaimBuster API [21] kullanılmıs¸tır. Kullandıkları veri seti, gelis¸tirdikleri bir veri toplama platformu ile kitle kaynaklı olarak etiketlenmis¸tir.
- **BERT:** BERT tabanlı modellerin çes¸itli NLP görevlerinde en gelis¸mis¸ modellerden daha iyi performans gösterdiği belirtildiği için, modelimiz sadece BERT kullanımına göre karşılaştırılmaktadır. İlgili eğitim veri setini kullanarak BERT modeline hassas ayar yapılmıs¸ ve hassas ayarlı model kullanılarak iddiaların teyit gerekliliği tahmin edilmiştir. Hem M-BERT hem de Temel BERT modeli (sadece ˙Ingilizce) için sonuçlar belirtilmis¸tir.
- **XLNET:** XLNet'in çeşitli NLP görevlerinde BERT'den üstün olduğu bildirilmektedir [40]. Bu nedenle, ilgili eğitim veri setiyle hassas ayar yapılarak çıktıları alınmıs¸ ve modelimiz ile kars¸ılas¸tırılmıs¸tır.
- **CTL'18, CTL'19 ve CTL'20-ED'nin En ˙Iyileri:** Her veri kümesi için, paylaşıldığı görevlerdeki en üst sıradaki katılımcılar olan sırasıyla Prize de Fer [42], Kopenhag [20] ve NLPIR@UNE [26] sonuçları da kars¸ılas¸tırılmıs¸tır.

4.2.1.3 Değerlendirme Ölçütleri

CTL görevlerinde kullanılan aynı değerlendirme yöntemi takip edilmektedir. Her dosya (tartışma veya konuşma dokümanı) için AP, RP ve P@10 değerleri hesaplanmış ve ortalamaları alınarak değerlendirmede kullanılmıştır.

4.2.2 Deney Sonuçları

Bu bölümde, araştırma sorusu için deneysel sonuçlar sunulmaktadır. İlk olarak, önerilen modelimizi çeşitli makine öğrenimi algoritmaları ile değerlendirilmiştir (Bölüm 4.2.2.1). Ardından, en iyi performans gösteren algoritmayı seçilerek ve her bir özniteliğin etkisini analiz etmek için özniteliklerin değerlendirilmesi çalışması yapılmıştır (Bölüm 4.2.2.2). Son olarak, yöntemlerimiz referans modeller ile karşılaştırılmıştır (Bölüm 4.2.2.3).

4.2.2.1 Öğrenme Algoritmalarının Karşılaştırılması

İlk deney kurulumunda, Bölüm 4.1'te tanımlanan tüm öznitelikler kullanılarak lojistik regresyon (LR), SVM, rastgele orman (RF), MART ve LambdaMART modellerini değerlendirilmiştir. Çizelge 4.2.2, her modelin MAP puanlarını göstermektedir. LR, tüm veri kümelerinde diğer tüm modellerden daha başarılı sonuçlar vermektedir. Lespagnol vd. [13] CTL'18 üzerinde yaptıkları benzer bir deneyde de, LR'ı kullandıklarında diğer modellerden daha yüksek sonuçlar elde ettiklerini bildirmişlerdir. Bu yüzden, sonraki deneylerde LR kullanılmaktadır.

Çizelge 4.2.2: Tüm Öznitelikleri Kullanan Farklı Modellerin MAP Puanları

Öğrenme Modeli	CTL18	CTL19	CTL20-ED
LR	**.2275**	**.1813**	**.0668**
RF	.1730	.1538	.0464
SVM	.1317	.1341	.0211
MART	.1781	.1753	.0246
Lambda MART	.0682	.0572	.0139

4.2.2.2 Özniteliklerin De ğerlendirilmesi

Kullanılan özniteliklerin etkinliğini analiz etmek için iki teknik uygulanmaktadır: 1) Bir tür öznitelik grubunu hariç tuttuğumuz ve modelin performansını onsuz hesapladığımız "bir öznitelik olmadan" metodolojisi ve 2) yalnızca tek bir öznitelik grubunun kullanıldığı "tek öznitelik" metodolojisi. Sonuçlar Çizelge 4.2.3'te gösterilmektedir.

Çizelge 4.2.3: Farklı Öznitelik Grupları için MAP Puanları

Bir Öznitelik Olmadan				**Tek Öznitelik**			
Öznitelik	**CTL 18**	**CTL 19**	**CTL 20-ED**	**Öznitelik**	**CTL 18**	**CTL 19**	**CTL 20-ED**
Hepsi	.2275	**.1813**	.0668				
Hepsi-MBERT	.2173	.1605	**.0989**	**MBERT**	.1850	**.1710**	.0231
Hepsi-KV	.1741	.1797	.0563	**KV**	**.2111**	.1428	**.0726**
Hepsi-TK	.2165	.1793	.0627	**TK**	.1361	.1050	.0610
Hepsi-POS	.2276	.1777	.0653	**POS**	.1047	.0635	.0673
Hepsi-KI˙	.2249	.1774	.0667	**KI˙**	.0796	.0675	.0290
Hepsi-KL	.2129	.1744	.0680	**KL**	.1538	.1098	.0314
Hepsi-YZ	**.2557**	.1771	.0654	**YZ**	.1007	.0593	.0338

Çizelge 4.2.3'teki sonuçlar, özniteliklerin her veri seti üzerinde farklı etkileri olduğu göstermektedir. MBERT, CTL'19'daki en etkili özniteliktir. Ancak, beklentilerimizin aksine KV, CTL'18'de MBERT'ten daha etkili bir öznitelik gibi sonuçlar

vermis¸tir. Ayrıca MBERT, CTL’20-ED’de en kötü sonuçları vermektedir. KV hariç tutulduğunda, tüm test koleksiyonlarında daha düşük performans elde edilmiştir. Ayrıca, yalnızca tek öznitelik kullanılan deneylerde CTL’18 ve CTL’20-ED’de KV ile en yüksek MAP puanına ulas¸ılmıs¸tır. CTL’19’da, yalnızca KV kullana- rak 0,1428 MAP puanı elde edilmiş ve bu, o koleksiyonda MBERT hariç diğer özniteliklerden daha etkili olduğunu göstermektedir.

Sonuçlarda, tartışmalı konuların (TK) da etkili özniteliklerden olduğu görülmektedir. Bunlar hariç tutulduğunda ise, tüm koleksiyonlarda modelin performansı biraz düs¸erken, yalnızca TK özniteliklerini kullanmak CTL’18 ve CTL’20-ED’de yüksek puanlar vermektedir ve CTL’18’deki en iyi performans gösteren katılımcıdan biraz daha iyi performans göstermektedir (Çizelge 4.2.3’te 0.1361’e kars¸ı 0.1332).

POS özniteliğinin hariç tutulması, CTL’19 ve CTL’20-ED’de modelin performansını biraz düşürürken, CTL’18’de neredeyse değiştirmemiştir. Ayrıca, Kİ’yi hariç tutmak, tüm koleksiyonlardaki performansı biraz düs¸ürmektedir. Benzer s¸ekilde, özel kelime listesi (KL) özniteliklerinin hariç tutulması, CTL’18 ve CTL’19’da performans düs¸üs¸üne neden olmaktadır. Yalnızca KL özelliklerini kullanmak ise, CTL’18’in tüm katılımcılarından daha iyi sonuçlar vermektedir (Çizelge 4.2.3’te 0.1538’e kars¸ı 0.1332).

Elde edilen bas¸arılı sonuçlar, bu listeyi genis¸letmenin daha fazla performans artışına yol açabileceğini göstermektedir. Yüklem zaman öznitelikleri ile ilgili olarak, sonuçlar ise karışıktır. Yüzlem zaman kipleri özniteliğinin hariç tutulması, CTL’19 ve CTL’20-ED’de hafif bir performans düs¸üs¸üne neden olurken, CTL’19’da yalnızca YZ kullanılması, diğer "yalnız bir kullanım" deneylerinden daha düşük sonuçlar vermektedir. Ayrıca, YZ hariç tutulduğunda CTL’18’de en iyi sonuç elde edilmektedir. Genel olarak, sonuçlar, YZ dışında tanımladığında tüm özniteliklerin, teyit gereklilik görevi için etkili olduğunu göstermektedir.

Çizelge 4.2.4: Rakip Modeller ile Kars¸ılas¸tırma. Rakip modeli kendimiz uygulayarak aldığımız sonuçlar, * işareti ile belirtilmiştir. En iyi sonuçlar **koyu** gösterilmis¸tir.

Model	CTL'18			CTL'19			CTL'20-ED		
	MAP	**RP**	**P@10**	**MAP**	**RP**	**P@10**	**MAP**	**RP**	**P@10**
ClaimBuster	.200	.216	.243	.133	.156	.200	.064	.039	.055
Lespagnol et al. [25]	.230	.254	.286*	.129*	.135*	.200*	.029*	.029*	.025*
Prise de Fer Team	.133	.135	.143	-	-	-	-	-	-
Copenhagen Team	-	-	-	.166	**.418**	.229	-	-	-
NLP&IR @UNE [26]	-	-	-	-	-	-	**.087**	**.093**	**.095**
XLNET	.197	.239	.257	.093	.077	.114	.028	.007	.015
M-BERT	.185	.222	.286	.170	.195	**.243**	.022	.021	.015
BERT	.179	.219	.229	.145	.211	.214	.061	.062	.065
Modelimiz	**.256**	**.266**	**.357**	**.177**	.208	.229	.065	.080	.080

4.2.2.3 Referans Modeller ile Kıyaslama

Ortalama olarak en yüksek MAP puanını elde ettiği için YZ dışındaki tüm öznitelikleri içeren model birincil model olarak seçilmis¸tir. İlk olarak, birincil modeli, eğitim için kullanılan koleksiyonlar ile ilgili referans modellerle karşılaştırma yapılmıştır (ClaimBuster farklı bir eğitim kümesinde önceden eğitilmiştir). Sonuçlar Çizelge 4.2.4'te sunulmus¸tur.

Önerilen modelimiz, CTL'18 koleksionunda diğer tüm modellerden tüm değerlendirme ölçütlerinde daha iyi performans göstermektedir. CTL'19'da ise CTL'de kullanılan resmi metrik olan MAP puanında en yüksek değere ulaşılmıştır. MBERT modeli, CTL'19'da P@10 puanı bakımından diğer modellerden daha iyi performans göstermektedir. RP değerlendirmesinde ise, Kopenhag Takımı en yüksek puanı almıştır. CTL'20-ED'de, tüm modellerin performansları diğer koleksiyonlara kıyasla daha azdır, hatta en iyi model yalnızca 0,087 MAP puanına ulas¸mıs¸tır. Bunun nedeni, CTL'20-ED'nin eğitim ve test kümelerindeki son derece düşük teyit gerektiren veri oranı olabilir (Bkz. Çizelge 4.2.1). Bununla birlikte, MBERT,

Lespagnol vd. ve XLNET, bu koleksiyonda düs¸ük performans gösterirken, NLPIR@UNE diğerlerinden belirgin bir şekilde daha iyi performansa sahiptir. Modelimiz ise, en iyi ikinci sonucu elde etmiştir. Çizelge 4.2.3'te gösterildiği gibi, bu koleksiyonda MBERT kullanımının modelimizin performansı üzerinde olumsuz etkisi vardır. MBERT dıs¸ındaki tüm öznitelikleri kullanan modelimiz 0,0989'a ulas¸arak NLPIR@UNE'den daha bas¸arılı olmus¸tur.

4.3. Nitel Analiz

Bu bölümde, ilgili koleksiyonun yalnızca eğitim verileriyle eğitilmiş modelimizin çıktıları (yüklem zaman özniteliği hariç tüm öznitelikleri kullanan lojistik regresyon) üzerine nitel analiz sunulmaktadır. Her girdi dosyası için, iddialar teyit gerekliliklerine göre sıralanmıs¸ ve ardından en yüksek puana sahip, teyit gerektirmeyen iddialar tespit edilmis¸tir. Çizelge 4.3.1 ve 4.3.2, CTL'18, CTL'19 ve CTL'20-ED test kümelerindeki tüm dosyalar için bu teyit gerektirmeyen ifadeleri göstermektedir. Nitel analizde aşağıdaki konuları belirlenmiştir.

- **Gelecekle İlgili İddialar:** 1. ve 34. satırlardaki ifadeler, geleceğe yönelik iddialardır. Yüklem zamanı özniteliğini içeren modelimiz, bu ifadeleri daha düs¸ük sıralarda sıralayabilir, ancak birincil modelimiz, ortalama olarak daha düşük performans sağladığı için bu özniteliği kullanmamaktadır.
- **Karmas¸ık Cümle Yapısı:** 2. ve 33. satırlardaki ifadelerin uzun ve karmas¸ık olması, belki de BERT modelinin ve ifadeleri temsil eden kelime vektörü özniteliklerinin performansını azaltmaktadır.
- **Veri Kümelerinin Etkileşimli Söylem Yapısı:** Kullandığımız veri kümeleri, siyasi tartıs¸maları ve konus¸maları kapsamaktadır. Bu nedenle, birçok ifade düzgün yapıda cümleler değildir veya konuşma sırasında kullanılan birçok eksik cümleler ve ünlem ifadeleri içerebilmektedir. Modelimizin, bu ifadeleri sıralamada yetersiz kaldığını görülmektedir. 3. ve 24-27 arası satırlardaki ifadelere bakıldığında, modelimizin bariz bir hata yaptığı ve eksik

cümlelere çok yüksek puanlar verdiği görülmektedir. Belki de bunun nedeni, bu ifadelerin önemli konular veya sayısal değerler hakkında kelimeler içermeleridir. Örneğin "jobs" kelimesi üçüncü satırdaki ifadenin, tanımladığımız tartışmalı konulardan biri olan işsizlikle ilgili olduğunu göstermektedir.

Ayrıca, belirli bir konuyla ilgili sayısal değerler, öznel açıklamalardan daha kolay doğruluğu kontrol edilebildiğinden, sayısal ifadeler sık sık teyit gerektiren iddialarda kullanılmaktadır. 28. ve 29. satırlardaki cümleler, daha önce ilgili tartıs¸mada bas¸ka bir kis¸i tarafından kesilen cümlelerin devamı gibi görünmektedir. Yine bu eksik cümleler, El Kaide ve uyus¸turucu gibi birçok önemli kelimeyi içermektedir. 17. satırdaki cümleyi çözümlemek zordur çünkü bu cümle gramer açısından düzgün değildir. 30. sıradaki ifadede ise, konus¸macı konus¸urken kendini düzeltmektedir.

Kelimeleri sıralaması, modelin tahminlerini etkilediğinden, bu tür cümlelerin dönüs¸türücü modellerinin performansını düs¸ürmesi de muhtemeldir. Fakat MBERT, genellikle gramer açısından doğru ve düzenli cümlelere sahip Wikipedia makaleleri üzerinde önceden eğitilmiştir. Bu nedenle MBERT, Satır 17 ve 30'daki gibi cümleler için iyi performans göstermeyebilmektedir.

- **Teyit Gerekliliğinin Öznelliği:** Vasileva vd. [35], doğruluk kontrolü yapan kurulus¸ların, seçilen iddialar arasından çok azı ile farklı iddiaları aras¸tırdıklarını belirtmis¸lerdir. Ayrıca TrClaim-19 analizlerinde (Bölüm 3.4) belirtildiği gibi, etiketleyicilerin iddiaların teyit gerekliliği konusunda sıklıkla fikir ayrılığına düştüğü de bildirilmiştir. Bu çalışmanın bulgularını doğrulayacak şekilde, 4-19. satırlardaki ifadelerin etiketlerinin öznelliği gözlemlenebilmektedir. Çünkü bunlar aslında gerçeklere dayalı iddialardır ve teyit gerektiren olarak kabul edilebilirler. Örneğin, 8., 11., 13., 15-19. satırlardaki ifadeler insanların oy verme kararını etkileyebilir. Bu nedenle, doğrulukları teyit gerektiren olabilir.

- **Genel İddialar:** 20. satırdaki iddianın doğruluğunu teyit etmek zordur çünkü iddia çok geneldir. Yoksulluğun artıp artmadığının kontrol edilebilmesi için,

referans zaman noktasına da ihtiyaç vardır. Bu nedenle, etiketleyicilerbunu teyit gerektirmeyen olarak değerlendirmiş olabilirler.

- **Zamirler:** 21-23. satırlardaki iddialar, ilgili tartıs¸mada önceki ifadelerde belirtilen belirli bir konuya veya kis¸iye atıfta bulunan zamirleri kullanmaktadırlar. Bu durum, modellerimizde bağlamın dikkate alınmasının performansı iyileştirebileceğini göstermektedir.

- **Kos¸ullu Cümleler:** 31. satırdaki ifade, eksik bir kos¸ullu cümledir. Kos¸ul cümlesindeki ifade ("Rusya hacklendi") teyit gerektiren olabilir. Bununla birlikte, koşul cümlesi nedeniyle gerçekte doğrulanması gereken bir iddia bulunmamaktadır. Görünüs¸e göre, modelimiz bunu yakalayamamaktadır. 32. satırdaki ifade, tam bir kos¸ullu cümle içermektedir. Cümlenin karmas¸ıklığı nedeniyle, kelime vektörleri ve MBERT öznitelikleri, iddianın gerçekte bir iddia olmadığını anlayamayabilir.

- **Sorunlu Etiketler:** Veri kümesindeki etiketler içinde tutarsızlıklar olduğu da görülmektedir. Örneğin, 9. satırdaki ("400.000 iş çıkardık") ifadesi farklı bir dokümanda (20160311_12_gop) da mevcuttur ve o dokümanda "teyit gerektiren" olarak eklenmis¸tir. Ek olarak, farklı etiketlere sahip anlamsal olarak çok benzer ifadeler de mevcuttur. Örneğin, 20160926_1_pres dokümanının 1079. satırındaki Donald Trump'ın "Irak'taki savas¸ı desteklemedim" ifadesi "teyit gerektirmeyen" olarak etiketlenirken, aynı dosyanın 1086. satırındaki "Irak'taki savas¸a kars¸ıydım" ifadesi, "teyit gerektiren "olarak etiketlenmiştir. Her iki ifade de benzer anlamlara sahiptir ve aynı bağlamda bulunmaktadır (dosyadaki konumları çok yakındır). Bu nedenle, her ikisi de aynı etiketlere sahip olabilir.

  Kars¸ı argüman olarak ise, "kars¸ı olmak"ın bir eylemi belirtmesi, "desteklememek"in herhangi bir eylemi belirtmemesi söylenebilir. Bu nedenle, benzer ifadeler için farklı etiketler, yine teyit gereklilik kararlarının öznelliğinden kaynaklanıyor olabilir.

Etiketine kesinlikle katılmadığımız ifadeler de mevcuttur. Örneğin, dokümanlardan birinde (20170315_nashville) Donald Trump'ın "Otomobil endüstrimizi tekrar işe koyacağız" ifadesi teyit gerektiren olarak etiketlenmiştir. Ancak ifade gelecekle ilgilidir ve doğrulanması mümkün değildir.

Yapılan nitel analizler, iddiaların teyit gerekliliğine dair etiketlemenin öznel olduğunu ve etiketlerin bazen yanlış olabileceğini göstermektedir. Kutlu vd. [24], belgelerdeki metin alıntılarının gerekçeler olarak kullanılmasının, ilgililik değerlendirmesindeki anlaşmazlıkları anlamaya yardımcı olduğunu göstermişlerdir. Bu fikri ile, TrClaim-19 olus¸turulurken teyit gerektiren etiketleri için gerekçeler de toplanmıs¸tır. Teyit gerektiren etiketlerinin arkasındaki gerekçeler, etiketin bir insan yargılama hatasından mı yoksa etiketleme görevinin öznelliğinden mi kaynaklandığının anlaşılmasına yardımcı olmaktadır. Ayrıca, bu etiketlerin arkasındaki gerekçeler, bu zorlu sorun için etkili çözümler geliştirilmesine katkı sağlamaktadır.

Çizelge 4.3.1: CTL'18 ve CTL'19 Test Dokümanlarında En Üstte Sıralanan Teyit Gerektirmeyen İfadeler. İfadelerin orijinali İngilizce olup Türkçe'ye çevrilmiş tir.

	Satır	İfade
CTL'1	1	CLINTON: Sahip olduğu plan bize mal olacak ve muhtemelen başka bir Büyük Buhran'a yol açacak.
	2	CLINTON: Sonra, New York Daily News röportajında, Sandy Hook ailesinin gençlere savas¸silahı, savas¸alanında öldürme reklamları yapılan AR-15'in reklamını dizginlemek için bir şeyler yapmaya çalışması için dava açıp desteklemeyeceği sorulduğunda, bunu iki katına çıkardı.
	3	TRUMP: Is¸ler, is¸ler, is¸ler.
	4	TRUMP: Bundan önce Demokrat Bas¸kan John F. Kennedy, ekonomiyi canlandıran ve is¸sizliği büyük ölçüde azaltan vergi indirimlerini savundu.
	5	TRUMP: Dünyanın en büyük s¸irketi Apple, Amerika'ya yurtdıs¸ından 245 milyar dolar kâr getirmeyi planladığını duyurdu.
	6	TRUMP: Amerika, Bill ve Hillary Clinton tarafından desteklenen feci ticaret anlas¸malarının yürürlüğe girmesinin ardından 1997'den beri imalat işlerinin neredeyse üçte birini kaybetti.
	7	TRUMP: Geçen yıl dünya ile ticaret açığımız yaklaşık 800 milyar dolardı.
CTL'19	8	O'MALLEY: Eğitim finansmanını yüzde 37 artırdık.
	9	KASICH: 400.000 is¸ kazandık.
	10	TAPPER: Eles¸tirmenler, bu anlas¸maların kurumsal Amerika'nın kar hanesi için harika olduğunu, ancak ABD'ye en az 1 milyon işe mal olduğunu söylüyor.
	11	TRUMP: İs¸sizlik bas¸vuruları 45 yılın en düs¸ük seviyesine ulas¸tı.
	12	TRUMP: Eğer düşünürseniz, şu ana kadar çelik dökümüne % 25, alüminyum dampingine % 10 gümrük vergisi koydum.
	13	TRUMP: Engelli Amerikalılar için is¸sizlik de tüm zamanların en düs¸ük seviyesine ulas¸tı.
	14	TRUMP: Tarihlerinde gördükleri en fazla sayıda cinayete sahipler - yaklas¸ık 40.000 cinayet.

Çizelge 4.3.2: CTL’20-ED Test Dokümanlarında En Üstte Sıralanan Teyit Gerektirmeyen ̇Ifadeler. Tweetlerin orijinali ̇Ingilizce olup Türkçe’ye çevrilmis¸tir.

CTL’20-ED	15	CRUZ: Ama bunların hepsi, 20 yıl önce Kuzey Kore ile yaptırımları kaldıran, milyarlarca doların akmasına izin veren ve bu parayı ilk etapta nükleer silah gelis¸tirmek için kullanan Clinton yönetiminin bas¸arısızlıklarının sonucudur.
	16	CRUZ: Yakın zamanda yapılan bir araştırma, Bernie’nin az önce söylediği Obamacare yetkilerinin arttığını buldu - en pahalı üç genç insan için yetki primleri yüzde 44 oranında arttı.
	17	TRUMP: Ama daha iyisini yapmalıyız çünkü Çin ile olan ticari açığımız, bildiğiniz gibi 504 milyar dolar.
	18	TRUMP: Demokratların sağlık hizmetlerini yok etme planı, yasadışı göçmenlerin yardımlarını finanse etmek için Medicare’e baskın yapmayı da içeriyor.
	19	TRUMP: Geçen yıl yaratılan yeni işlerin % 60’ını kadınlar doldurdu ve kadın işsizliği şu anda 74 yılın en düs¸ük seviyesinde ..
	20	SANDERS: Yoksulluk artıyor.
	21	TRUMP: Özel bir e-postası vardı.
	22	TRUMP: Rusya’yı temsil ediyor.
	23	TRUMP: Düs¸ünülemez - 1,8 milyar dolar.
	24	TRUMP: Kırk dört milyar.
	25	TRUMP: Yüzde yirmi bes¸.
	26	TRUMP: Rusya, Rusya, Rusya, Rusya.
	27	BIDEN: Anneler - örgüt - s¸iddete kars¸ı anneler - silahlı s¸iddet.
	28	CLINTON: ... Afganistan-Pakistan tiyatrosunda El Kaide ile mücadelede.
	29	BUTTIGIEG: ... veya reçeteli ilaçların maliyeti.
	30	TRUMP: Kanada ile büyük bir ticaret açığımız var, okudum, oh, aslında bu fazlalık.
	31	TRUMP: Rusya hacklenirse, Rusya bizim seçimimizle bir ilgisi varsa,
	32	TRUMP: Ama bir ticaret anlas¸masında bize Kuzey Kore konusunda yardım etselerdi, yardım etmemelerinden kesinlikle çok daha kolay olurdu.
	33	RON DESANTIS: Pinellas County’de körfezin hemen karşısında büyüyüp saatliği altı dolardan çalıs¸maya bas¸layan biri olarak burada bulunmak, Amerika Birles¸ik Devletleri Bas¸kanı tarafından onaylanmıs¸ olmak gerçek bir onurdur.
	34	DE BLASIO: Ben başkan olduğumda, skoru eşitleyeceğiz ve burayı daha adil bir ülke yapmak ve çalışan insanları ilk sıraya koyan bir ülke olmasını sağlamak için zenginlerden vergi alacağız.

5.E ğitim Veri Kümesinin Etkisi

Önceki bölümde (Bölüm 4) önerilen modelimizde, MBERT ile farklı öznitelikler kullanılmaktadır. Ancak, önceki çalıs¸malar sadece dönüs¸türücü modellerini hassas ayarlamanın bile teyit gereklilik ve diğer NLP görevlerinde çok başarılı sonuçlar ortaya çıkarabileceğini göstermiştir. Bu nedenle, AS-5: İnce ayarlanmış BERT modelinin teyit gereklilik görevi için daha fazla etiketli veri kullanması performansı artıtır mı? ve AS-6: Eğitim verisi nasıl artırılabilir ? araştırma sorularına odaklanılmıs¸tır.

5.1.E ğitim Veri Kümesi Artırma Yöntemleri

Bu bölümde, BERT modellerine hassas ayar yapmak için elle etiketlenen veri boyutunu artırılabilecek dört farklı yöntem açıklanmaktadır.

5.1.1 Rastgele Artırma (RastArt)

Bu yöntemde "Herhangi bir ek etiketli veri dönüs¸türücü modellerinin performansını artırır mı?" sorusuna cevap aramak için, etiketli veriler rastgele seçilip eğitim verilerine eklenmis¸tir.

5.1.2 Teyit Gerektiren İddia Sayısını Artırma (TGArt)

Mevcut veri kümelerinin çoğunda teyit gerektiren iddiaların sayısı çok azdır (Bkz. Çizelge 4.2.1). Bu nedenle, dönüşümcü modelleri, teyit gerektiren iddiaların gerçek özelliklerini tam olarak öğrenemeyebilir. Ayrıca, negatif örnekler, teyit gerektirmeyen iddialar, eklemek, etiket dağılım dengesizliğini artırarak potansiyel olarak modellerin performansını düşürebilmektedir. Bu nedenle, bu yöntemde, etiket dengesizliğini azaltmak için pozitif örnekler, teyit gerektiren iddialar, eklenmiştir.

5.1.3 Aktif Öğrenme ile Artırma (AÖArt)

Rastgele seçilen veriler, modeli etkili bir şekilde eğitmek için yararlı olmayabilir. Bu nedenle, bu bölümde etiketlenecek verileri seçmek için aktif bir öğrenme yöntemi uygulanmaktadır. Model öncelikle temel veri kümesinde eğitilir ve eğitilmiş model kullanılarak diğer veriler sıralanır. Daha sonra, en üst sıradaki N iddia seçilerek önceki eğitim kümesine eklenir. En üst sıradaki iddiaları seçmek, teyit gerektirebilecek iddiaların seçilmesine olanak sağlamaktadır. Bu yüzden, etiket dağılımınındaki dengesizlik de azaltılabilir.

5.1.4 Çok Dilli Eğitme

Yanlış bilgi küresel bir sorun olsa da, her dil için ayrı bir model oluşturmak son derece zordur. Çok dilli modeller bu sorun için etkili bir çözüm olabilir. Aynı zamanda, her dil için iddiaları etiketlemek hala maliyetli bir süreçtir. Bu nedenle, bu yöntemde, diller arası eğitimin teyit gereklilik görevi için ne kadar etkili olduğunu araştırılmıştır. Özellikle, çeşitli dillerdeki etiketli veriler kullanıla- rak MBERT modeline hassas ayar yapılmış ve belirli bir dilde değerlendirilmiştir. Örneğin, CTL İngilizce veri kümeleri kullanılarak MBERT modeline hassas ayar yapılarak, Türkçe için teyit gereklilik veri kümesi olan TrClaim-19 üzerinde test edilmiştir. Mevcut veri kümeleri İngilizce, Arapça ve Türkçe olduğu için, bu diller üzerinde çalışma yapılmıştır.

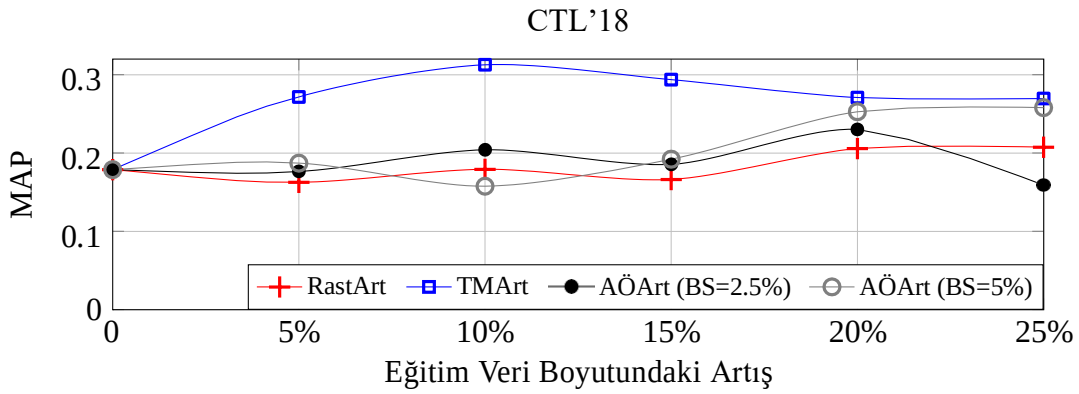
5.2. Deneyler

5.2.1 Deney Sonuçları

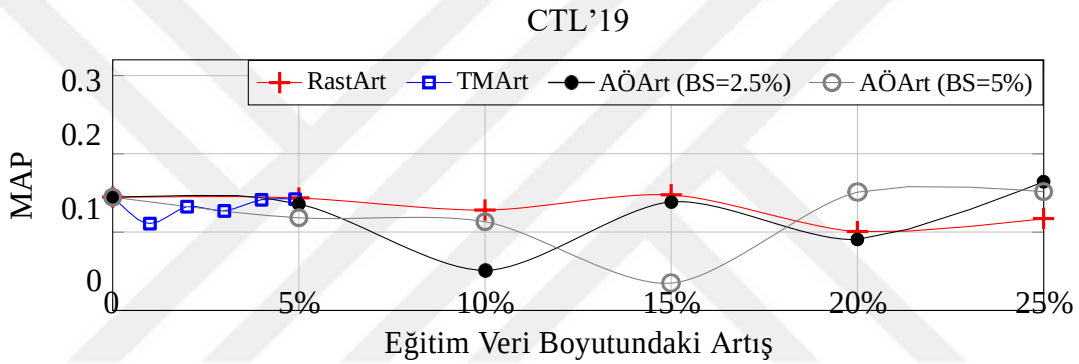
Bu bölümde, eğitim verisi boyutunun BERT modellerinin(AS-5) performansı üzerindeki etkisini analiz etmek, eğitim veri boyutunu (AS-6) artırmak için en iyi yöntemi bulmak ve çok dilli eğitimin performans üzerindeki etkisini gözlemlemek için deneyler yapılmış¸tır. Özellikle BERT kullanılmış¸tır çünkü CTL'20 görevlerinin birçok katılımcısı çes¸itli dönüs¸türücü modelleriyle etkileyici sonuçlar alınabileceğini göstermiştir. Ayrıca, MBERT modelini kullanarak çok dilli eğitim gerçekleştirilebilirken, daha önce belirtilen öznitelikler bu deneylerde doğrudan uygulanamamaktadır.

5.2.1.1 Tek Dilli E ğitim

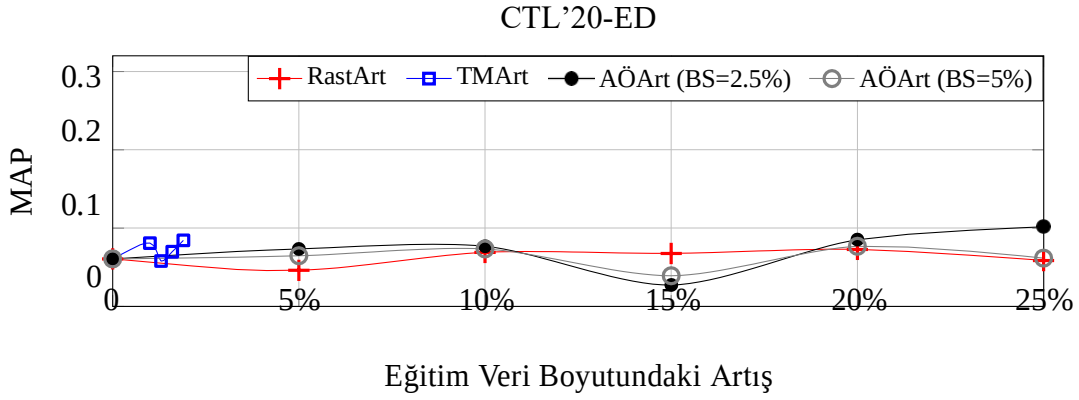
İlk olarak Bölüm 5.1'te açıklanan RastArt, TGArt ve AÖArt tekniklerini kullanarak CTL'18, CTL'19 ve CTL'20-ED eğitim verileri artırılarak elde edilen veri kümeleriyle Temel BERT modeli hassas ayar yapılmış¸tır. Ardından, her bir koleksiyonun test verileri üzerindeki hassas ayarlı BERT modellerinin performansı değerlendirilmiştir. Belirli bir koleksiyonla yapılan deneylerde, eğitim veri boyutunu artırmak için diğer tüm İngilizce verileri kullanılmıştır. Örneğin, yöntemleri değerlendirmek için CTL'18 kullanıldığında, CTL'19, CTL'20-ED ve CTL'20-ET'deki tüm verileri kullanılmıştır. Eğitim veri boyutu, her koleksiyonun eğitim veri boyutunun % 25'ine kadar artırılmaktadır.



(a)



(b)



(c)

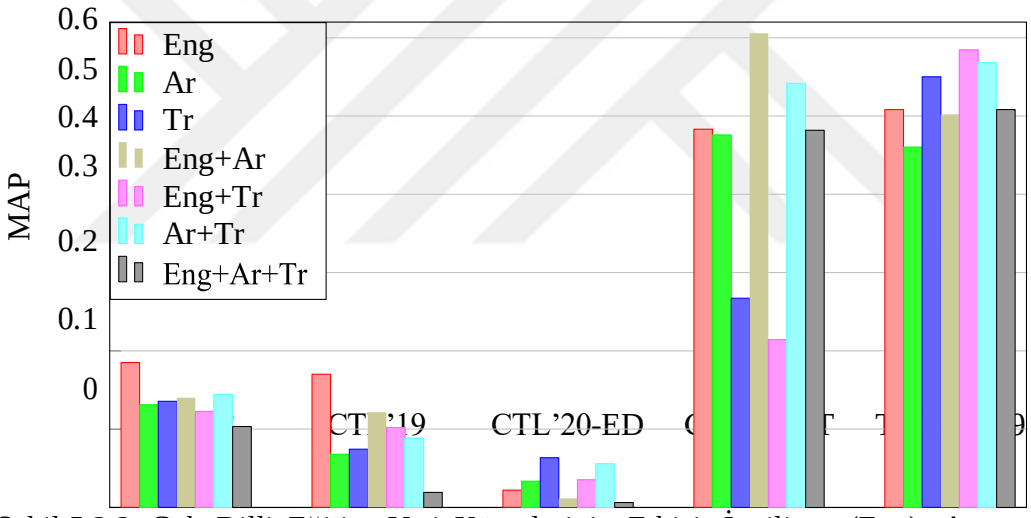
Şekil 5.2.1: Eğitim Verisinin Artırmanın Etkisi. "BS", artırılan veri boyutunu (batch size) ifade etmektedir.

Her koleksiyonun farklı bir boyutu vardır (Bkz. Çizelge 4.2.1) ve bu yüzden deneyler için gereken veri miktarı değişkendir. Bu nedenle, CTL'19 ve CTL'20-ED ile yapılan deneylerde, yetersiz sayıdaki teyit gerektiren iddia nedeniyle eğitim veri boyutu TGArt ile % 25'e kadar artırılamamaktadır. Bu nedenle, CTL'19 ve CTL'20-ED'de TGArt yöntemi kullanılarak eğitim veri boyutu sırasıyla % 4,9 ve % 1,9 artırılmış,tır. AÖArt yönteminin performansı iki farklı eklenen veri grubu boyutu için ölçülmüş,tür. Özellikle, AÖArt yönteminin eklenen veri grubu boyutu, ilgili koleksiyonun eğitim veri boyutunun % 2,5'i ve % 5'i olarak ayarlanmıştır. TGArt ve RastArt için, verilerin rastgele seçilmesi nedeniyle deneyler üç kez tekrarlanarak sonuçların ortalaması alınmıştır. Sonuçlar Şekil 5.2.1'de gösterilmektedir.

Rastgele iddia seçerek eğitim veri boyutu artırıldığında, BERT modellerinin performansı doğrusal olarak artmamaktadır. Aslında, CTL'19 ve CTL'20-ED veri kümelerinde eğitim verileri RastArt yöntemi kullanılarak % 25 artırıldığında BERT modelinin performansı düşmektedir. Bu nedenle, sonuçlara bakıldığında, daha fazla etiketlenmis¸ verinin, teyit gereklilik görevi için hassas ayarlanmıs¸ BERT modellerini kullanırken her zaman daha yüksek performans sağlamadığı görülmektedir.

TGArt, yalnızca teyit gerektiren iddiaların eklenmesi, tüm CTL'18 deneylerine diğerlerinden daha başarılı sonuçlar vermektedir. Özellikle, TGArt kullanılarak eğitim veri boyutu % 10 artırıldığında, CTL'18 için şu ana kadar bildirilen en yüksek MAP puanı olan 0,3127 elde edilmis¸tir. Ayrıca, TGArt kullanılarak CTL'20-ED veri kümesinde veri kümesi boyutu yalnızca % 1,9 artırıldığında, neredeyse diğer tüm denemelerden daha yüksek sonuçlar elde edilmiştir. Ancak TGArt, CTL'19'da BERT modelinin performansını iyiles¸tirmemis¸tir. TGArt için sonuçlar ümit verici olmasına rağmen eşleşmeyen artış oranları nedeniyle TGArt çalışmasını diğerleriyle tam olarak karşılaştırmak mümkün değildir. Ancak, daha fazla etiketli veriye sahip olunduğunda, TGArt yöntemi daha uygulanabilir olacaktır. Ayrıca, teyit gerektirmeyen verilerin etiketleme maliyeti nedeniyle TGArt'ı uygulamanın maliyeti de çok daha yüksek olacaktır.

AÖArt ile ilgili gözlemler ise burada belirtilmis¸tir. Öncelikle, farklı eklenen veri grubu büyüklüklerine sahip AÖArt yöntemleri arasında büyük bir performans farkı vardır. Bu nedenle, eklenen veri grubu boyutunu optimize etmek için daha fazla deney gereklidir. Bu aras¸tırmayı gelecekteki bir çalıs¸ma olarak yapılabilir. İkinci olarak, CTL'19'da her iki AÖArt yönteminin performansında bir salınım gözlemlenmektedir. Bunun nedeni, MAP puanını hesaplamak için yalnızca yedi tartıs¸ma metni kullanılması olabilir. Tek bir metin için AP puanındaki herhangi bir değişikliğin MAP puanı üzerinde büyük etkisi olabilmektedir. Üçüncüsü, TGArt ve AÖArt'yı karşılaştıran sonuçlar da karışıktır. CTL'18 ve CTL'20-ED'deki çoğu durumda, AÖArt, RastArt'tan daha yüksek MAP puanları vermektedir. Ancak TGArt, CTL'19 denemelerinin çoğunda AÖArt'den daha başarılıdır. AÖArt, ortalama olarak RastArt ile veri artırmadan biraz daha iyi performans göstermektedir.



Şekil 5.2.2: Çok Dilli Eğitim Veri Kümelerinin Etkisi. İngilizce (Eng), Arapça (Ar) ve Türkçe (Tr) iddialar kullanılarak hassas ayar yapılan MBERT modelinin performansı gösterilmis¸tir. "Eng" ifadesi, CTL'18,CTL'19 ve CTL'20-ED ile yapılan deneylerde, kendi eğitim kümelerini ifade ederken; CTL'20-AT ve TrClaim-19 deneylerinde CTL'20-ET eğitim kümesi için kullanılmıs¸tır.

5.2.1.2 Çok Dilli E ğitim

Bir sonraki deney düzeneğinde, farklı dillerdeki verileri kullanarak MBERT modellerine hassas ayar yapılmış, ve İngilizce (CTL'18, CTL'19, CTL'20-ED), Arapça (CTL'20-AT) ve Türkçe (TrClaim-19) koleksiyonlar üzerindeki performansları değerlendirilmiştir. İngilizce koleksiyonlar ile ilgili deneylerde, TrClaim-19, CTL' 20-AT ve ilgili İngilizce koleksiyonunun eğitim verilerinin çeşitli kombinasyonları kullanılmış,tır. Benzer s,ekilde, Arapça ve Türkçe koleksiyonlarla deneyler yapılırken, farklı TrClaim-19, CTL'20-AT ve CTL'20-ET kombinasyonları kullanılmış,tır. Arapça ve Türkçe veri kümeleri üzerinde deneyler yaparken, İngilizce iddialar için diğer İngilizce koleksiyonlar yerine CTL'20-ET kullanılmıştır. Çünkü CTL'20- ET, CTL'20-AT ve TrClaim-19, tweetlerden oluşurken diğerleri politik tartışmalar ve konuşmalar içermektedir. Sonuçlar Şekil 5.2.2'de gösterilmektedir.

İngilizce olmayan verilerin kullanılmasının CTL'18 ve CTL'19 koleksiyonlarında modelin performansını düşürdüğünü gözlemlenmiştir. Ayrıca, tüm dilleri kullanmak, tüm İngilizce koleksiyonlarında en düş,ük performansa neden olmaktadır. CTL'18 ve CTL'19 koleksiyonlarında en iyi performans yalnızca İngilizce veriler kullanıldığında elde edilmektedir. İlginç bir şekilde, TrClaim-19 ile hassas ayar yapılan model CTL'20-ED'de en iyi sonuçları vermektedir. Yalnızca Arapça verilerin kullanılması da CTL'20-ED'de yalnızca İngilizce veriler kullanılmasından daha bas,arılı sonuçlar vermektedir.

CTL'20-AT ile yapılan deneylerde, İngilizce verilerle hassas ayar yapılan modelin, Arapça verilerle yapılan hassas ayarlı modelden biraz daha iyi performans göstermektedir. İngilizce ve Arapça verileri birlikte kullanıldığında en iyi sonuçları elde edilmektedir. Ayrıca, Türkçe ve Arapça verileri kullanmak ikinci en iyi sonuçları verirken üç dilin birden kullanıldığı durumdan daha iyi performans göstermektedir.

Türkçe veri kümesi üzerinde yapılan deneylerde tek dil kullanıldığında en iyi performansı Türkçe ile eğitilen model göstermektedir. Yalnızca Arapça veriler ile hassas ayar yapıldığında Türkçe için en kötü sonuçlar elde edilmiştir. Türkçe eğitim kümesi, Arapça ya da İngilizce veri kümesi ile genişletildiğinde modelin ba-

şarısı az da olsa artmaktadır. En iyi sonuçlar ise İngilizce veri ile genişletildiğinde elde edilmektedir. Ancak, üç dili de kullanarak elde edilen performans, MBERT modellerinde hassas ayar yapmak için yalnızca İngilizce veya İngilizce ve Arapça verileri kullanıldığında elde edilen performansa benzemektedir.

Genel olarak, İngilizce olmayan verilerin kullanılmasının çoğu durumda İngilizce iddialar için performansı düşürdüğünü gözlemlenmektedir. Öte yandan Arapça ve Türkçe iddialarda, iddiaların orijinal diline ek olarak başka bir dilde veri kullanılması modellerin performansını artırmaktadır. Bununla birlikte, her üç dil de kullanıldığında, diller arası eğitimin olumlu etkisinin azaldığı görülmüştür.

5.2.1.3 Modelimiz ve BERT'in Artırılmış Eğitim Verileri ile Karşılaştırılması

Bölüm 5.1'te açıklanan yöntemler kullanılarak ek eğitim verilerinin kullanıldığı önerilen modelimiz, BERT modeliyle karşılaştırılmıştır. Şekil 5.2.1'de gösterilen önceki deneylerde kullanılan en büyük eğitim veri kümelerinde önerilen model eğitilmiş ve BERT hassas ayar yapılmıştır. AÖArt ve RastArt için ilgili koleksiyonun % 25'i kadar eğitim verileri artırılmıştır. Ancak, TGArt için, sınırlı sayıdaki teyit gerektiren iddianın olması nedeniyle eğitim verileri CTL'18, CTL'19 ve CTL'20-ED için sırasıyla % 25, % 4.9 ve % 1.9 artırılmıştır. RastArt ve TGArt yöntemleri üç kez tekrarlanmış ve önceki deneylerde olduğu gibi ortalama performansı belirtilmiştir. Sonuçlar Çizelge 5.2.1'te gösterilmektedir.

RastArt yöntemi kullanıldığında, modelimiz her durumda BERT'den daha iyi performans göstermektedir. Eğitim verileri, TGArt tarafından artırıldığında, modelimiz CTL'18 ve CTL'20-ED'de BERT'den daha iyi performans gösterirken; BERT, CTL'19'da daha yüksek MAP ve RP puanlarını elde etmiştir. Eğitim verilerini artırmak için AÖArt kullanıldığında ise, modelimiz CTL'20-ED'de BERT'den daha başarılı olmasına rağmen BERT, diğer koleksiyonlardaki denemelerin çoğunda daha yüksek sonuçlara ulaşmıştır.

Eğitim veri kümesi artırma yöntemleri karşılaştırıldığında, çoğu durumda TGArt en iyi sonuçları vermektedir. TGArt'ı kullanarak, CTL'20-ED için şu ana kadar bildirilen en yüksek MAP puanına (0.117) ulaşılmıştır. Sonuçlar, genel olarak, MBERT modelini çeşitli özniteliklerle birleştiren modelimizin çoğu durumda BERT modelinden daha başarılı olduğunu göstermektedir.

Çizelge 5.2.1: Farklı Yöntemlerle Artırılan Eğitim Verisinin BERT(B) ve Modelimizin(M) Karşılaştırılması. AÖArt ve RastArt yöntemlerinde, ilgili eğitim kümeleri, boyutlarının %25'i kadar artırılmıştır. TGArt yönteminde ise, kısıtlı teyit gerektiren iddia sayısından dolayı CTL'18, CTL'19 ve CTL'20-ED eğitim kümeleri sırasıyla %25, %4.9 ve %1.9 artırılmıştır. En iyi sonuçlar **koyu** gösterilmiştir.

Yöntem		**CTL'18**			**CTL'19**			**CTL'20-ED**		
		MAP	**RP**	**P@10**	**MAP**	**RP**	**P@10**	**MAP**	**RP**	**P@10**
RastArt	**B**	.205	.220	.224	.120	.141	.138	.060	.039	.045
	M	.234	.263	.324	.148	.178	.195	.099	.081	.077
TGArt	**B**	.264	.261	.338	.145	.155	.186	.086	.081	.085
	M	**.282**	**.303**	**.390**	.141	.145	.214	**.117**	**.125**	**.110**
AÖArt	**B**	.254	.263	.329	**.153**	.166	**.243**	.067	.045	.050
	M	.236	.252	.286	.143	**.182**	.200	.098	.072	.075

6. SONUÇ VE ÖNERİLER

Bu tez çalışmasında, Türkçe için ilk etiketli teyit gerektiren iddia veri kümesi olan TrClaim-19 sunulmuş, teyit gerektiren iddiaların önceliklendirilmesi için hassas ayarlanmış BERT ve farklı öznitelikler kullanan güdümlü bir öğrenme modeli önerilmiş ve eğitim veri kümesini artırmanın BERT ve önerilen model üzerindeki etkileri incelenmiştir.

TrClaim-19'un oluşturulması için öncelikle Türkiye'deki önemli olaylarla ilgili 225 milyon Türkçe tweet toplanmıştır. Toplanan tweetlerden, etiketlenecek olanları seçmek için iki Türkçe doğruluk kontrolü yapan web sitelerinden doğruluğu kontrol edilmiş 765 iddia edinilmiştir. Her bir iddia için 3 tweet olmak üzere toplamda 2287 adet tweet etiketlenmek için seçilmiştir. Teyit gereklilik etiketlerine ek olarak, teyit gerektiren etiketlerinin ardındaki gerekçeler de toplanmıştır. Bu gerekçeler üzerine yapılan analizler sonucu, iddiaların teyit gerekliliği kararında etiketleyicilerin birbirine fazla katılmadığı, uzmanlar ve uzman olmayanların çoğu kararda hemfikir olmadığı, iddiaların teyit gerektiren olabilmesi için çok farklı gerekçeler olabileceği, iddianın konusunun teyit gerekliliğini etkileyen en önemli etken olduğu ve olumsuz olayların teyit gerekliliğini olumlu olaylardan daha fazla etkilediği gözlemlenmiştir.

Teyit gerektiren iddiaların önceliklendirilmesi için önerdiğimiz model, hassas ayar yapılmış, BERT ve yerel tartış,malı konular, kelime vektörleri, özel kelime listesi, kars,ılas,tırma ifadeleri, POS etiketleri ile yüklem zaman kiplerini öznitelik olarak kullanmaktadır. Yüklem zaman kipleri hariç, kullanılan tüm özniteliklerin teyit gereklilik görevinde etkili olduğu gözlemlenmiştir. Ancak, özniteliklerin etkililiği veri kümesine göre değişmektedir. Örneğin, MBERT iki veri kümesinde en başarılı öznitelik iken, diğerinde en başarısız olanıdır. Önerilen modelimiz, CTL'18 ve CTL'19 koleksiyonlarında en iyi modellerden daha bas,arılı olmus,tur.

Eğitim veri kümesini artırmak için rastgele veri artırma, teyit gerektiren veri sayısını artırma, aktif öğrenme ile veri artırma ve çok dilli veri kümeleri oluşturulması yöntemleri kullanılmıştır. Eğitim veri boyutunu artırmanın her zaman performansı artırmadığı gözlemlenmiştir. Hem BERT'in hem de önerilen yöntemimizin başarısı en etkili artıran yöntem eğitim veri kümesini sadece teyit gerektiren etiketli verilerle genis,letmektir. Bunun sebebi teyit gerektiren etiketli verilerin veri kümelerinde çok az bulunması olabilir. Çok dilli eğitim ise Arapça ve Türkçe veri kümelerinde performansı artırırken İngilizce için etkili olmamıştır. Eğitim veri kümesi artırılarak modelimiz uygulandığında CTL'20-ED koleksiyonunda en bas,arılı sonuçlar elde edilmis,tir.

Gelecekte, zayıf denetim teknikleri ve makine tercümesi ile bas,ka sözcüklerle ifade etme gibi otomatik veri artırma yöntemleri üzerinde çalıs,malar yapmak planlanmaktadır. Yapılan nitel analize göre, teyit gerektiren verilerin gerekçelerinin belirtilmesi faydalı olacaktır. Gerekçelerin yararlı olarak kullanılması da çalıs,ılması gereken alandır. Çalışma yapılacak bir diğer alan ise, teyit gerektiren iddiaların önceliklendirilmesinde açıklamalı yapay zeka yöntemlerinin kullanılmasıdır. Bu sayede, bir iddianın neden teyit gerektiren olduğu daha iyi anlaşılabilir. Böyle bir modele sahip olmak, gerçek hayatta daha uygulanabilir olup kullanıcıları uyararak yanlıs, bilginin yayılmasını azaltacaktır.

Kaynakça

[1]A GEZ, R., BOSC, C., LESPAGNOL, C., PETITCOL, N., AND MOTHE, J. IRIT at checkthat! 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018* (2018).

[2]A LTUN, B., AND KUTLU, M. TOBB-ETU at CLEF 2019: Prioritizing claims based on check-worthiness. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019* (2019).

[3]A TANASOVA, P., NAKOV, P., KARADZHOV, G., MOHTARAMI, M., AND DA SAN MARTINO, G. Overview of the clef-2019 checkthat! lab on automatic identification and verification of claims. task 1: Check-worthiness. In *CEUR Workshop Proceedings* (2019).

[4]B ARRÓN-CEDENO, A., ELSAYED, T., NAKOV, P., DA SAN MARTINO, G., HASANAIN, M., SUWAILEH, R., HAOUARI, F., BABULKOV, N., HAMDAN, B., NIKOLOV, A., ET AL. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (2020), Springer, pp. 215–236.

[5]B ARRÓN-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. S. Overview of checkthat! 2020: Automatic identification and verification of claims in social me-

dia. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (Cham, 2020), Springer International Publishing, pp. 215–236.

[6]B ARRÓN-CEDEÑO, A., ELSAYED, T., NAKOV, P., MARTINO, G. D. S., HASANAIN, M., SUWAILEH, R., AND HAOUARI, F. Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media. *Advances in Information Retrieval 12036* (2020), 499 – 507.

[7]C HERUBINI, F., AND GRAVES, L. The rise of fact-checking sites in europe. *Reuters Institute for the Study of Journalism, University of Oxford* (2016).

[8]C OCA, L. G., CUSMULIUC, C., AND IFTENE, A. Checkthat! 2019 UAICS. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019* (2019).

[9]D EVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (2019), pp. 4171–4186.

[10]D HAR, R., DUTTA, S., AND DAS, D. A hybrid model to rank sentences for check-worthiness. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019* (2019).

[11]F AVANO, L., CARMAN, M. J., AND LANZI, P. L. Theearthisflat's submission to clef'19checkthat! challenge. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019* (2019).

[12]F LEISS, J., ET AL. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin 76*, 5 (1971), 378–382.

[13]F LETCHER, R., CORNIA, A., GRAVES, L., AND NIELSEN, R. K. Measuring the reach of "fake news" and online disinformation in europe. *Reuters Institute Factsheet* (2018).

[14]F RIEDMAN, J. Greedy function approximation: A gradient boosting machine. *Annals of Statistics 29* (2001), 1189–1232.

[15]G ASIOR, J., AND PRZYBYLA, P. The ipipan team participation in the check-worthiness task of the clef2019 checkthat! lab. In *CLEF (Working Notes)* (2019).

[16]G ENCHEVA, P., NAKOV, P., MÀRQUEZ, L., BARRÓN-CEDEÑO, A., AND KOYCHEV, I. 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* (2017), pp. 267–276.

[17]G HANEM, B., MONTES-Y-GÓMEZ, M., PARDO, F. M. R., AND ROSSO, P. UPV-INAOE - check that: Preliminary approach for checking worthiness of claims. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018* (2018).

[18]H ANSEN, C., HANSEN, C., SIMONSEN, J. G., AND LIOMA, C. 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 checkthat! lab. In *CLEF* (2018).

[19]H ANSEN, C., HANSEN, C., SIMONSEN, J. G., AND LIOMA, C. Neural weakly supervised fact check-worthiness detection with contrastive sampling-based ranking loss. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019* (2019).

[20]H ANSEN, C., HANSEN, C., SIMONSEN, J. G., AND LIOMA, C. Neural weakly supervised fact check-worthiness detection with contrastive sampling-based ranking loss. In *CLEF* (2019).

[21]H ASSAN, 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. Claimbuster: The first-ever end- to-end fact-checking system. *PVLDB 10* (2017), 1945–1948.

[22]J ARADAT, I., GENCHEVA, P., BARRÓN-CEDEÑO, A., MÀRQUEZ, L., AND NAKOV, P. 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* (2018), pp. 26–30.

[23]K ARTAL, Y. S., AND KUTLU, M. Tobb etu at checkthat! 2020: Prioritizing english and arabic claims based on check-worthiness. *Cappellato et al.[10]* (2020).

[24]K UTLU, M., MCDONNELL, T., BARKALLAH, Y., ELSAYED, T., AND LEASE, M. Crowd vs. expert: What can relevance judgment rationales teach us about assessor disagreement? In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (2018), ACM, pp. 805–814.

[25]L ESPAGNOL, C., MOTHE, J., AND ULLAH, M. Z. Information nutritional label and word embedding to estimate information check-worthiness. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2019), ACM, pp. 941–944.

[26]M ARTINEZ-RICO, J., ARAUJO, L., AND MARTINEZ-ROMO, J. Nlp&ir@ uned at checkthat! 2020: A preliminary approach for check-worthiness and claim retrieval tasks using neural networks and graphs.

[27]M CDONALD, T., DONG, Z., ZHANG, Y., HAMPSON, R., YOUNG, J., CAO, Q., LEIDNER, J., AND STEVENSON, M. The university of sheffield at checkthat! 2020: Claim identification and verification on twitter. *Cappellato et al.[10]*.

[28]M IKOLOV, T., CHEN, K., CORRADO, G., AND DEAN, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[29]M ORSTATTER, F., SHAO, Y., GALSTYAN, A., AND KARUNASEKERA, S. From alt-right to alt-rechts: Twitter analysis of the 2017 german federal elec-

tion. In *Companion Proceedings of the The Web Conference 2018* (2018), pp. 621–628.

[30]NAKOV, P., BARRÓN-CEDEÑO, A., ELSAYED, T., SUWAILEH, R., MÀRQUEZ, L., ZAGHOUANI, W., ATANASOVA, P., KYUCHUKOV, S., AND DA SAN MARTINO, G. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (2018), pp. 372–387.

[31]PATWARI, A., GOLDWASSER, D., AND BAGCHI, S. 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* (2017), ACM, pp. 2259–2262.

[32]PIRES, T., SCHLINGER, E., AND GARRETTE, D. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), pp. 4996–5001.

[33]SCHWETER, S. Berturk - bert models for turkish, Apr. 2020.

[34]SU, T., MACDONALD, C., AND OUNIS, I. Entity detection for check-worthiness prediction: Glasgow terrier at CLEF checkthat! 2019. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019* (2019).

[35]VASILEVA, S., ATANASOVA, P., MÀRQUEZ, L., BARRÓN-CEDEÑO, A., AND NAKOV, P. 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* (2019).

[36]VOORHEES, E. M. The philosophy of information retrieval evaluation. In *Workshop of the Cross-Language Evaluation Forum for European Languages* (2001), Springer, pp. 355–370.

[37]VOSOUGHI, S., ROY, D., AND ARAL, S. The spread of true and false news online. *Science 359*, 6380 (2018), 1146–1151.

[38]WILLIAMS, E., RODRIGUES, P., AND NOVAK, V. Accenture at checkthat! 2020: If you say so: Post-hoc fact-checking of claims using transformer- based models. *arXiv preprint arXiv:2009.02431* (2020).

[39]WU, Q., BURGES, C. J., SVORE, K. M., AND GAO, J. Adapting boosting for information retrieval measures. *Inf. Retr. 13*, 3 (June 2010), 254–270.

[40]YANG, Z., DAI, Z., YANG, Y., CARBONELL, J., SALAKHUTDINOV, R. R., AND LE, Q. V. Xlnet: Generalized autoregressive pretraining for langu- age understanding. In *Advances in Neural Information Processing Systems* (2019), pp. 5754–5764.

[41]YASSER, K., KUTLU, M., AND ELSAYED, T. bigir at CLEF 2018: Detection and verification of check-worthy political claims. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum* (2018).

[42]ZUO, C., KARAKAS, A., AND BANERJEE, R. A hybrid recognition system for check-worthy claims using heuristics and supervised learning. In *CLEF* (2018).