TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

AVRUPA BİRLİĞİ KAMU ALIMI İHALELERİ İLAN METİNLERİNİN EKONOMİK ETKİLERİNİN DOĞAL DİL İŞLEME İLE İNCELENMESİ

YÜKSEK LİSANS TEZİ

Mustafa Kaan GÖRGÜN

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Öğr. Üyesi Mücahid KUTLU

HAZİRAN 2022

TEZ BİLDİRİMİ

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

Mustafa Kaan GÖRGÜN

İMZA

ÖZET

Yüksek Lisans Tezi

AVRUPA BİRLİĞİ KAMU ALIMI İHALELERİ İLAN METİNLERİNİN EKONOMİK ETKİLERİNİN DOĞAL DİL İŞLEME İLE İNCELENMESİ

Mustafa Kaan GÖRGÜN

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

Tez Danışmanı: Dr. Öğr. Üyesi Mücahid KUTLU

Tarih: HAZİRAN 2022

Küresel bazda kamu alımları 11 trilyon dolar tutarı ile muazzam bir ekonomik aktiviteye denk gelmektedir. 250 binden fazla Avrupa Birliği kamu otoritesi her yıl Avrupa Birliği'nin (AB) gayri safi yurtiçi hasılasının (GSYİH) yüzde 14'ünü oluşturan yaklaşık 2.3 trilyon euroyu kamu alımları için harcamaktadır. Yapılacak küçük çaplı iyileştirmeler bile kamu kaynaklarının kullanımını büyük ölçüde etkilemektedir. Bu nedenle kamu kaynaklarının etkin kullanımı için rekabetçi ve maliyet etkin kamu alım süreçlerinin oluşturulması önem arz etmektedir. Kamu otoriteleri ise piyasa rekabet gücünün sağlanması, verilerin şeffaflığı ve güvenilirliği gibi Avrupa Birliği Komisyonunun direktiflerininin uygulamasından sorumludur. Kamu alımları faaliyetlerini kontrol edebilmek için ilan edilmeden önceki süreçte ihale ölçütlerini tahmin edebilmek önem arz etmektedir. Binlerce belge içerisinden ihale ölçütlerinin belirlenmesi başlı başına önemli bir problem olarak karşımıza çıkmaktadır. Bu çalışmada kullanılan veri kümesi, 2011'den 2018'e kadar 22 AB üye ülke dilini kapsayan AB'nin Resmi Gazetesine Ekinin çevrimiçi versiyonu olan Tenders Electronic Daily'deki (TED) kamuya açık sözleşme belgelerindeki sözleşme açıklama metinlerinden derlenmiştir. Gerçek veriler kullanılarak yapılan deneylerde, çok dilli hassas ayarlanan dönüştürücü modeller, cümle temsil vektörleri tabanlı yaklaşımlar ve öznitelik tabanlı yaklaşımlar önerilen metotlar olarak sunulmaktadır. Önerilen metotlar yalnızca ilan açıklamalarını kullanarak tek teklifli sözleşmeler, toplam teklif sayısı, yabancı firma sözleşmeleri ve sözleşme fiyatının etkinliği olmak üzere dört ekonomik ölçütü tahmin etmek için kullanılmaktadır.

Önerilen yöntemler, tüm tahmin görevlerinde kelime frekans vektörlerini ve ekonomik göstergeleri kullanan modeller de dahil olmak üzere tüm temel modellerden daha iyi performans göstererek, ilan açıklamalarının kamu ihalelerinin sonuçlarında önemli bir rol oynadığını göstermektedir. Ayrıca, çok dilli eğitimin tüm görevlerde orjinal dildeki eğitime göre tahmin performansını geliştirdiği görülmektedir. Bu ise kullanılan veri kümesinin diğer ülkelerin kamu ihale ilanları için de kullanılabileceğini göstermektedir. Geliştirilen modeller, kamu ihale yetkilileri tarafından ihale ilanlarının inceleme sürecini otomatikleştirmek ve düşük rekabete neden olanları tespit etmek için kullanılabilir. Ayrıca katılımcı firmalar karşılaşacakları potansiyel rekabeti tahmin etmek için geliştirilen modelleri kullanarak daha öngörülebilir kararlar alabilir ve geleceğe yönelik risklerini azaltabilirler. Aynı zamanda belirtilmelidir ki geliştirilen modeller sadece kamu alımı ilan metinlerindeki açıklamaları kullanmaktadır. Bu nedenle model performansları ihale süreçlerini etkileyen diğer ekonomik etkenler dahil edilerek iyileştirmeye açık olmakla birlikte bu çalışmanın kapsamının dışında kalmaktadır.

Anahtar Kelimeler: Kamu alımları, Çok dilli metin analitiği, Uygulamalı makine öğrenimi

ABSTRACT

Master of Science

INVESTIGATION OF THE ECONOMIC IMPACT OF THE EUROPEAN UNION PUBLIC PROCUREMENT NOTICE DESCRIPTIONS BY USING NATURAL LANGUAGE PROCESSING

Mustafa Kaan GÖRGÜN

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

Supervisor: Dr. Öğr. Üyesi Mücahid KUTLU

Date: JUN 2022

Global public procurement amounts to $11 trillion annually. More than 250,000 European Union (EU) public authorities spend around €2.3 trillion every year which constitutes 14 percent of the European Union GDP. Even small improvements in the procurement processes have a large effect on the effectiveness of public funds usage. Therefore, competitive and cost-effective public procurement processes are essential for the effective use of public resources. Public authorities in EU are responsible for applying EU commission directives such as ensuring market competitiveness, transparency and reliable data. To control public procurement activities, predicting future award metrics before bidding is crucial while there exist thousands of documents about procurement calls. The data used in this work consists of European Union's multilingual public procurement notices. The data covers raw contract description texts from 2011 to 2018 spanning 22 EU languages and extracted from publicly available contract documents in Tenders Electronic Daily (TED) which is the online version of 'Supplement to the Official Journal' of the EU. In the experiments, fine-tuned multilingual transformer models, sentence embeddings approach, and a feature based approach are proposed to predict four economic metrics including the single bid awards, the number of offers, foreign contract awards, and contract price effectiveness. In all prediction tasks, proposed methods outperform all baselines including models that use word ngrams and economic indicators. This suggests that notice descriptions play an important role in the outcome of public procurement calls.

We also showed that multilingual training improves prediction accuracy in all tasks suggesting that EU dataset can be also used for public procurement calls of other countries. Developed models can be used by public procurement officials to automatize the examining process of procurement notices and detect the ones causing low competition. Participating firms can also use the models to predict potential competition they will face to make better decisions and reduce the future risks. Note that there is still gap to improve model performances by adding economic indicators into models. But the best performing models in this work focus only procurement notice descriptions.

Keywords: Public procurement, Multilingual text analytics, Applied machine learning

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren hocalarım Dr. Öğr. Üyesi Mücahid KUTLU'ya ve Prof. Dr. Bedri Kamil Onur TAŞ'a kıymetli tecrübelerinden faydalandığım TOBB Ekonomi ve Teknoloji Üniversitesi Bilgisayar Mühendisliği Bölümü öğretim üyelerine ve en önemlisi destekleriyle her zaman yanımda olan aileme ve arkadaşlarıma teşekkür ederim.

Çalışmalarım boyunca ARDEB 1001 programının 119K986 numaralı projesi kapsamında araştırma bursu desteğiyle beni destekleyen TÜBİTAK'a da ayrıca teşekkür ederim.

İÇİNDEKİLER

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

KISALTMALAR

CPV : Common Public Vocabulary
GSYİH : Gayri Safi Yurt İçi Hasıla
ISO : International Organization for Standardization
NER : Named Entity Recognition
MAE : Mean Absolute Error
MBERT : Multilingual Bidirectional Encoder Representations from Transformers
POS : Part of Speech
TED : Tenders Electronic Daily
XLMR : Cross-lingual Language Model - Roberta

1. GİRİŞ

Kamu alımları, yetkili kamu kurum ve kuruluşları tarafından mal ve hizmetlerin satın alınmasını ifade etmektedir. Dünya Bankası raporlarına göre [2] [25], küresel gayri safi yurtiçi hasılanın (GSYİH) yüzde 12'sinin kamu alımları için harcandığı tahmin edilmektedir. Avrupa Birliği'nde (AB) ise, 250 binden fazla AB kamu otoritesi her yıl AB GSYİH'nın yüzde 14'ünü oluşturan yaklaşık 2.3 trilyon euroyu kamu alımları için harcamaktadır [1]. Bu durum göstermektedir ki, ekonomide bu ölçüde büyük bir paya sahip olan kamu alımları süreçlerinde yapılacak küçük çaplı iyileştirmeler bile kamu kaynaklarının kullanımını büyük ölçüde etkilemektedir. Kamu kaynaklarının etkili kullanımı ise vergi mükellefi olarak toplumun her bir bireyini ilgilendirdiği için kamu otoriterlerini büyük bir sorumluluk altına almaktadır. Dünya bankası raporuna göre [31] kamu alım sistemlerinin zayıf yönetimi, kamu yatırımlarını büyük siyasi ve ekonomik yükümlülüklere çevirebilir, kalkınma hedeflerini ve sonuçlarını engelleyebilir ve ek maliyetlere ve kamu fonlarının boşa harcanmasına neden olabilir. Avrupa Komisyonu'nun kamu alımlarına ilişkin direktifleri [10], kamu kaynaklarının etkili bir şekilde kullanılmasını sağlamak için, daha geniş bir erişim alanı oluşturulmasının yanında şeffaflığın ve güvenilir verilerin artırılması gibi stratejik ilkelerin korunmasını amaçlamaktadır. 2014/24 numaralı komisyon direktifinde teknik şartnamenin müteşebbislerin ihale prosedürüne eşit erişimini sağlayacak ve kamu ihalelerinin rekabete açılması önünde haksız engeller yaratmasını engelleyecek nitelikte olmasını zorunlu kılmaktadır. İlan edilen bu direktiflere uygun şekilde şeffaflığı teşvik etmek ve firma katılımını artırmak için, kamu kurum ve kuruluşları AB Resmi Gazetesine Ekinde (Supplement to the Official Journal) kamu alım duyurularını yayınlamaları gerekmektedir. AB Resmi Gazetesi Ek'nin çevrimiçi versiyonu olan "Tenders Electronic Daily" (TED)[2], haftada yaklaşık 2,400, yılda 746 binden fazla ihale duyurusu yayınlamaktadır. Bu sistemi kullanarak, ilgilenen firmalar AB üye ülkelerinde düzenlenen kamu alım ilanlarına göz atabilir ve kendi uzmanlık alanına veya bölgesine göre arama yapabilir.

Kamu alım ilanları düzenli ve umuma açık şekilde yayınlanmasına ve komisyon düzenlemelerinin yürürlükte olmasına rağmen ilan içerikleri büyük ölçüde farklılık göstermekte ve alım süreçleri göründüğünden daha karmaşık olabilmektedir. Bazı durumlarda, teknik ayrıntılar çok uzun olmakta ve daha fazla bilgiye ulaşmak ancak dış kaynaklara (resmi web siteleri vb.) yönlendiren bağlantılarla sağlanmaktadır. Bu da ayrı ayrı okunması ve incelenmesi gereken onlarca belgenin ortaya çıkmasına neden olmaktadır. Önemli miktarda kamu ilanı ise teknik şartname ve ihale süreçleri hakkında yeterli bilgi içermemektedir. Yeterli miktarda bilgi içermeyen ilanlar ise yabancı ülkelerde iş fırsatları arayan firmalar için ek zorluklar oluşmasına, rekabetin azalmasına ve dolayısıyla satın alma maliyetlerinin artmasına neden olmaktadır.

[1] https://simap.ted.europa.eu/en_GB/web/simap/european-public-procurement
[2] https://ted.europa.eu/TED/

Buna bağlı olarak kamu kurum ve kuruluşları ilanlardaki bilgi içeriklerini düzenleyerek ihale sürecinin işleyişini değiştirebilirler. Örneğin, başka firmaların ihaleye girişlerini engellemek için önemli teknik bilgileri saklayabilir veya sözleşme detaylarını anlaşılması zor bir şekilde yazabilirler. Ayrıca yerel dillerde çıkan ilanlar yabancı ve yerli firmalar arasında teknik detayları takip etmek açısından fırsat eşitsizliği oluşturup yerel firmaları daha avantajlı hale getirebilir. Gürcistan kamu ilanları üzerinde yapılan bir araştırmaya göre [5] hem Gürcü dilinde hem de İngilizce olarak ilan edilen kamu alımlarında sadece Gürcü dilinde yapılan ilanlara göre katılımın büyük oranda arttığı gözlenmiştir. Yayınlanan çok sayıda ilan dikkate alındığında, sınırlı sayıda kaynağa sahip AB düzenleyicilerinin, her bir kamu satın alma süreci için sürecin adil, rekabetçi ve uygun maliyetli olup olmadığını ayrı ayrı değerlendirmeleri mümkün değildir. Bu nedenle, kamu kaynaklarının etkin kullanımına yardımcı olmak için otomatik sistemlere ihtiyaç vardır.

Bu tez çalışmasında Avrupa Birliği kamu alım ilan açıklamalarını kullanarak rekabetin ve maliyet etkinliğinin metinlerden tahmin edilmesi amaçlanmıştır. Rekabet ve maliyet etkinliğini ölçmek için ise ekonomi literatüründe kullanılan [17, 21, 30] dört adet ölçüt kullanılmaktadır. Rekabet ölçütleri olarak teklif sayısı, ihaleye birden fazla teklifin yapılıp yapılmadığı ve yabancı bir firmanın ihaleyi kazanıp kazanmadığı bilgileri kullanılmaktadır. Maliyet etkinliği ölçütü olarak ise sözleşme fiyatının ilana çıkan kurum veya kuruluşun tahmin ettiği fiyatı aşıp aşmadığı değerlendirilmektedir.

Yapılan deneyler bahsi geçen dört ölçütü kullanarak üç sınıflandırma ve bir regresyon problemi olmak üzere toplamda dört problem olarak ele alınmıştır. İlk aşama deneylerde öznitelik çıkarımları yapılarak bu özniteliklerin belirlenen ölçütleri tahminlemedeki etkileri incelenmiştir. Ardından ilan açıklamalarının çok-dilli cümle temsil vektörleri kullanılarak bu cümle temsil vektörleri ile farklı makine öğrenmesi modelleri kullanılarak tahmin performansları belirlenmiştir. Ayrıca sektör bazında tahminlemenin performansa etkisini ölçmek amaçlı 45 sektör grubu için tahminler alınarak temel model performanslarıyla karşılaştırılmıştır. Temel modeller literatürde uygulanan kelime frekans vektörleri ve ekonomik göstergeler olmak üzere genel olarak iki grup öznitelikleri kullanmaktadırlar. Kullanılan veri kümesi 22 farklı Avrupa diline ait metinleri içermesi nedeniyle çok dilli dönüştürücü modellerin hassas ayar performansları da deneylerde incelenmiştir. Hem hassas ayarlanmış çok dilli dönüştürücü modellerde hem de çok dilli cümle temsil vektörlerinin oluşturulmasında önceden eğitilen MBERT [7] ve XLMR [3] modelleri kullanılmıştır. Yapılan deneyler sonucunda, MBERT hassas ayarlanmış dönüştürücü modeli tek teklif, teklif sayısı ve yabancı firma sözleşmesi tahmin problemlerinde bütün modeller arasında en iyi performansı sergilemektedir. MBERT cümle temsil vektörlerini girdi olarak kullanan MLP tabanlı yaklaşım ise maliyet etkinliğini tahmin probleminde temel model performanslarından daha iyi olduğu gözlemlenmiştir.

Bu çalışmada aşağıda listelenen araştırma sorularına cevaplar aranmıştır.

1. **Soru:** Kamu alım ilan metinleri ihale sürecinde rekabeti ve maliyet etkinliğini etkiliyor mu?
   Cevap: Sadece ilan metinlerini kullanan modeller sadece makroekonomik göstergeleri kullanan temel modellerden daha iyi başarım elde etmektedir. Bu ise ilan metinlerinin ihale süreçleri üzerinde etkili olduğunu göstermektedir.

2. **Soru:** Önerilen hangi yaklaşımlar hangi problemler üzerinde daha iyi sonuçlar vermektedir?
   Cevap: Hassas ayarlanan dönüştürücü modeller arasında MBERT, XLMR'a göre bütün problemlerde daha iyi sonucu vermektedir. Ayrıca maliyet etkinliği probleminde cümle temsil vektörleri ile eğitilen MLP modeli hassas ayarlanan modellerden daha iyi sonuçlar vermektedir.

3. **Soru:** Çoklu dilde yapılan model eğitimi dil özelinde yapılan model eğitimlerine göre nasıl performans sergilemektedir?
   Cevap: Yapılan deneyler göstermektedir ki hassas ayarlamada eğitim kümesindeki bütün dil örneklerini kullanan tek bir MBERT modeli, her bir dil için sadece o dile ait örnekleri kullanan farklı modellerin tahmin sonuçlarının birleştirilmesiyle elde edilen tahminlerden daha iyi sonuçlar vermektedir. Bu durum göstermektedir ki farklı dillerin kullanılmasıyla tahmin performansı geliştirilebilir.

4. **Soru:** Farklı dillerde yapılan model eğitimleri ve değerlendirmelerde hangi dil çiftleri daha iyi performans göstermektedir?
   Cevap: Hassas ayarlamada orjinal dilin kullanılması sadece teklif sayısını tahminlemede diğer kombinasyonlardan daha iyi performans göstermektedir. Bütün test kümesinde ise Tek Teklifli ihalelerin tespitinde orjinal dilin yanında İngilizce örneklerin hassas ayarlamada kullanılması performansı artırmaktadır. Fakat hassas ayarlamada sadece İngilizce örnekleri kullanmak performansı oldukça düşürmektedir.

5. **Soru:** Öznitelik tabanlı modellerde hangi öznitelikler ilgili ölçütleri tahminlemede daha başarılı?
   Cevap: Genel olarak öznitelik temelli model performansları diğer önerilen model performanslarına göre daha düşük kalmaktadır. Fakat sektör ve ülke bilgisinin teklif sayısı ve yabancı firma problemlerinde daha etkili olduğu tek teklif probleminde ise ihale metninin hangi dilde yazılmış olduğu bilgisinin önemli olduğu gözlenmiştir.

Yapılan tez çalışmasının literatüre katkısı aşağıda listelenmiş beş noktada toplanabilir.

1. AB kamu ilan alım metinlerinin ihale sürecine etkilerini araştıran literatürdeki ilk araştırma olma özelliği taşımaktadır.

2. Çok dilli dönüştürücü modellerin 22 farklı Avrupa dilindeki kamu ilan metinleri üzerinde dil geçişli analizleri literatüre kazandırılmıştır.

3. Deney sonuçları göstermektedir ki ilan metin içeriklerinin düzenlenmesi ihale süreçlerinin rekabetçi ve maliyet etkin olmasında etkili olmaktadır.

4. Geliştirilen modeller hem ihale politikasını düzenleyen AB görevlileri hem de katılımcı firmalar ve kamu kurum ve kuruluşları için yol gösterici olarak kullanılabilir.

5. Paylaşılan hassas ayarlanmış çok dilli dönüştürücü modellerin 100'den fazla dil için kullanılabilmesi sayesinde birçok ülkede ihale süreçlerinin düzenlenmesinde yardımcı olarak kullanılabilir.

Bu tez çalışmasının kalan kısmı şu şekilde düzenlenmiştir: **Bölüm 2**'de çalışma ile ilgili literatürde yapılan benzer çalışmalardan, **Bölüm 3**'te çalışmada kullanılan veri kümesinin oluşturulma aşamalarından ve veri kümesinin sahip olduğu genel özelliklerden, **Bölüm 4**'te tahmin ölçütlerinden ve problem tanımlarından, **Bölüm 5**'te önerilen ve temel yöntemlerden, **Bölüm 6**'da ise deney sonuçlarından ve sonuçlara bağlı değerlendirmelerden bahsedilmiş ve son olarak **Bölüm 7**'de yapılan çalışma kısaca özetlenmiş ve gelecekte yapılabilecek çalışmalardan söz edilmiştir.

2. LİTERATÜR ARAŞTIRMASI

Bu çalışma yapısı itibariyle hem ekonomi literatüründe hem de bilgisayar bilimlerinde olan çalışmalarla ilgilenmektedir. Ayrıca ihale dokümanları ile yapılan çalışmalar ile çok dilli metin sınıflandırma çalışmalarına da bu bölümde yer verilmiştir.

2.1 Ekonomi Literatüründeki Çalışmalar

Kamu alımlarının ekonomik önemi sebebiyle ekonomi disiplininde bu alanda birçok araştırma bulunmaktadır. Iimi [13], Japon resmi kalkınma yardımı projeleri için düzenlenen satın alma ihalelerinde teklif sayısındaki yüzde 1'lik bir artışın sözleşme fiyatını yüzde 0,2 oranında azalttığını bildirmektedir. Estache ve Iimi [8] çalışmasında teklif yapan firma sayısının artmasının rekabeti desteklediğini ve ihalelerin önemli ölçüde daha düşük fiyatlarla sonuçlandığını raporlamaktadır. Benzer şekilde, Onur vd. [22], Türk kamu ihalelerinde teklif sayısının artmasının sözleşme fiyatlarını büyük oranda azalttığını göstermektedir. Onur ve Taş [23] ise yine Türk kamu ihalelerinde 1'den 8 teklif sayısına kadar teklif sayısının artmasıyla sözleşme maliyetlerinin düşüş göstermekte olduğunu söylemektedir.

2.2 Bilgisayar Bilimlerindeki Çalışmalar

Bilgisayar bilimlerinde kamu alımlarındaki düşük teklif [33] sayısına bağlı maliyet artışı, projelerin teklif verme süreçlerindeki belirsizlik riskleri [15, 18], teklif veren sayısı [6, 27] ve sözleşme fiyatı [11] gibi konularda çalışmalar bulunmaktadır.

Demiray vd. [6] Türk kamu alım ihalelerindeki teklif sayısını, tahmin edilen sözleşme fiyatı, ihalenin mal veya hizmet ihalesi olup olmadığı ve yabancı firmalara açık olup olmadığını içeren 9 farklı öznitelik kullanarak SVM modeli ile tahmin etmektedirler. Ayrıca kullandıkları veri kümesi 85,052 satırdan oluşmaktadır. Teklif sayısını tahmin ederken teklifleri düşük orta ve yüksek olmak üzere 3 sınıfa ayırmaktadırlar. Yüksek teklif sayısı alan sınıfın en kötü sınıflandırılan sınıf olduğunu bildirmişlerdir.

Rabuzin ve Modrusan[27] ise Hırvat kamu alımlarında ilana çıkılan ihalelerin bir veya birden fazla teklif alıp almayacağını tahmin etmektedir. Tek teklif alan ihaleleri yolsuzluk açısından "şüpheli" olarak değerlendirmektedirler. Naive Bayes (NB), Logistic Regression (LR) ve SVM modellerini TF-IDF öznitelikleri kullanarak karşılaştırdıktan sonra en iyi performansı LR modeliyle aldıklarını raporlamaktadırlar.

Görgün vd. [12] bu çalışmada kullanılan aynı veri kümesinin yalnızca 2018 yılını içeren alt kümesini kullanarak teklif sayısını tahmin etmek için, adlandırılmış varlıklar, açıklama uzunlukları, TF-IDF vektörleri, sektör bilgisi, ülke ve dil kodlarını kullanmaktadırlar. Bi-gram frekans vektörlerini kullanarak KNN modelinin en yüksek

performansı verdiğini bildirmişlerdir. Çalışmalarında İngilizce olmayan tüm açıklamaları otomatik olarak İngilizce'ye çevirmekte ve öznitelik çıkarımlarını İngilizce metinler üzerinden yapmaktadırlar.

Garcia Rodriguez vd. [11] 2012-2018 arası İspanyol kamu alım ihalelerinde sözleşme fiyatını tahmin etmek için tarih, sektör kodu ve bölgeyi içeren 14 adet öznitelik kullanarak Random Forest Modeli geliştirmişlerdir. Kullandıkları veri kümesi 2012'den 2018'e kadar olan yıllardaki 58,337 adet ihale dokümanını kapsamaktadır. Ayrıca teklif sayılarını incelerken sadece bir teklif alan ihaleleri "rekabet olmayan", 2 ila 4 arası teklif alan ihaleleri "düşük teklifli", 5 ila 10 arası teklif alan ihaleleri "orta teklifli" ve daha yüksek teklif alan ihaleleri "yüksek teklifli" olarak sınıflandırmaktadırlar.

Lima vd. [19] ise çalışmalarında Brezilya kamu alım ilanlarından 15 milyon dokümanı kullanmışlardır. Bu dokümanlardan 1907 tanesi uzmanlar tarafından yolsuzluk ve dolandırıcılık açısından riskli olarak etiketlemiş ve bu riskli ilanları belirlemek için ise bi-LSTM modeli geliştirmişlerdir. Geliştirdikleri modelin lineer modellere göre az bir farkla daha iyi performans gösterdiğini kaydetmektedirler.

Bu tez çalışması ise buraya kadar olan Görgün vd. [12] haricindeki çalışmalardan farklılaşmaktadır. Bu çalışmada kullanılan veri kümesi birden çok sayıda ülkede ve birden çok sayıda dilde yapılan ilanları kapsamaktadır. Ayrıca bu çalışma dört alt problem olarak ele alınmakta, çok dilli dönüştürücü modellerin performansları analiz edilmekte ve çoklu dillerin performanslar üzerindeki etkileri incelenmektedir.

2.3 TED Açık Verisi Kullanılarak Yapılan Çalışmalar

TED verilerinin herkese açık bir kaynak olmasıyla bir takım araştırmacılar bu verileri kullanarak farklı çalışmalar yapmışlardır. Ahmia vd. [1] NLP çalışmaları, özellikle makine çevirisi alanı, için farklı dillerdeki karşılıklı cümlelerden oluşan bir veri kümesi oluşturmuşlardır. Simperl vd. [28] ise TED veri kümesinden bir bilgi çizgesi oluşturmakta ve bu çalışmada dokümanlar arası bağlantıların oluşturulmasında zorlayıcı özelliklerinden biri olarak da verinin çoklu dilde olmasını göstermektedirler. Mehrbod ve Grilo [20] ise TED verisi üzerinde semantik aramayı kolaylaştırmak için veri kümesindeki adlandırılmış varlıkları tespit etme üzerine çalışmışlardır.

2.4 Çoklu Dilde Metin Sınıflandırma Çalışmaları

Çoklu dilde örnekleri içermesiyle bu çalışma literatürdeki çoklu dildeki doküman sınıflandırma çalışmalarıyla da ilgilenmektedir. Pappas ve Popescu-Belis [24] çok dilli hiyerarşik dikkat ağları yaklaşımını önermekte ve yaklaşımlarını 8 ayrı dilde yazılan haber makaleleri üzerinde değerlendirmektedirler. van der Heijden vd. [32] ise çok dilli ve dil geçişli görevler için meta-öğrenme yaklaşımı önermektedir. Wu ve Dredze [34] ise MBERT modelini doküman sınıflandırma, cümle ögelerine ayırma, adlandırılmış varlıkları çıkarma, bağımlılık ayrıştırma ve doğal dil anlamlandırma gibi görevler üzerinde incelemektedir. MBERT modelinin her bir görevde en iyi sonucu verdiğini raporlamaktadırlar. Benzer şekilde Karthikeyan vd. [16] çalışmasında MBERT modelinin dil geçişlilik performansını, model mimarisini ve farklı öğrenme objektiflerini incelemektedirler. Sözlüksel yakınlık olan dillerin dil geçişli senaryolarda

ihmal edilebilir performans değişiklikleri gösterdiğini raporlamaktadırlar. Pires vd. [26] ise MBERT modelinin İngilizce'den farklı cümle yapısına sahip dillerdeki performansının düşük olduğunu göstermektedir. Bu çalışmada ise çok dilli model eğitimleri için MBERT ve XLMR modellerinden faydalanıyoruz. Literatüre bakıldığında bu çalışma 22 Avrupa dili için çok dilli yapılan ilk çalışma özelliğini taşımaktadır.

3. KULLANILAN VERİ KÜMESİ

Bu tez çalışmasında kullanılan veri kümesi Avrupa Birliği Yayın Ofisi'nin yayınladığı ihale dokümanlarının (public procurement notices) derlenmesiyle oluşturulmuştur. Yayınlanan dokümanlar herkese açık olarak Avrupa Birliği Açık Veri Portalı web sayfasından[3] ulaşılabilir. Ayrıca yayınlanan dokümanlardan çıkarılmış ve yapılandırılmış veri setleri de paylaşılmaktadır. Bu veri setleri sözleşme ilanları (contract notices) ve sözleşme sonuç ilanları (contract award notices) olarak iki grupta toplanmaktadır. Veri setlerinin içerdiği bazı alanlara örnek olarak sözleşme ilanları için ilan tarihleri, sözleşme tipi, ihaleye çıkan otoritenin ismi, şehri, ülkesi gibi bilgiler, ihalede talep edilen işin hangi sektörle alakalı olduğuna dair sektör kodları (CPV codes), eğer ihale lotlara ayrılıyorsa lot numaraları; sözleşme sonuç ilanları için ise kazanan firmanın ismi, adresi, ülkesi, posta kodu gibi bilgiler; ihale için alınan teklif sayısı (number of offers), ihalenin sonuç fiyatı (award price) verilebilir.

Bu tez çalışmasının odaklandığı sözleşmelerin açıklama kısımları ise yayınlanan veri setlerinde yer almamakla birlikte yayınlanan dokümanlarda mevcuttur. Bu sebeple kullanılacak açıklama kısımları dokümanlardan çekilip, yayınlanan veri setleri ile birleştirilmiş ve bu sayede hangi sözleşme metninin hangi sözleşme veya sözleşme sonuç bilgilerine sahip olduğu elde edilmiştir. Ardından yapılan ön işleme aşamasından sonra deneyler için kullanılacak veri kümesi elde edilmiştir. Deneylerde kullanılan veri kümesine ait istatistikler Çizelge 3.1'de görülebilir.

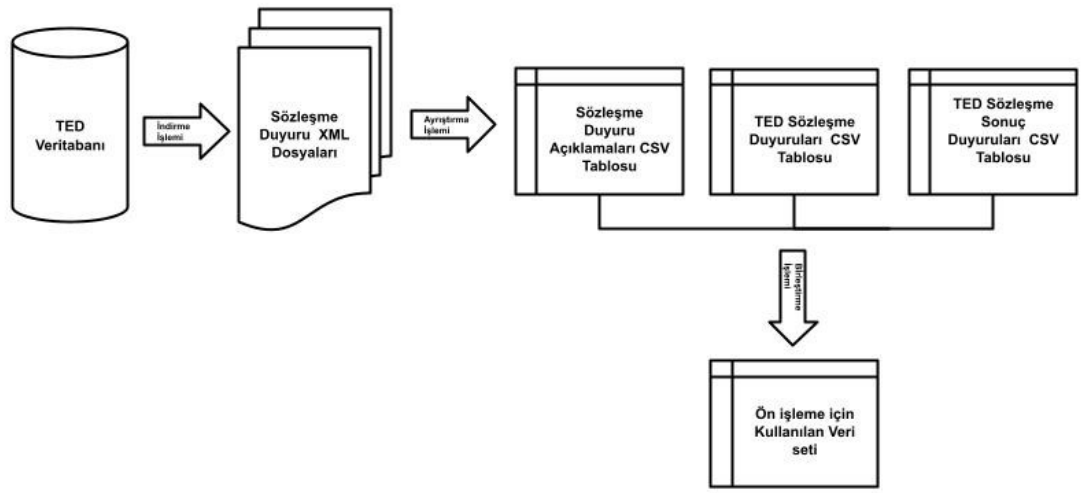
Çizelge 3.1: Deneylerde kullanılan veri kümesine ait istatistikler

Kapsanan Yıllar	2011-2018
Tek Teklif problemi için sözleşme sayıları	333832
Teklif Sayısı problemi için sözleşme sayıları	333832
Yabancı Firma problemi için sözleşme sayıları	333832
Fiyat Etkin problemi için sözleşme sayıları	106122
Tek teklif alan ihalelerin oranı	21.0%
Effektif olmayan ihaleleri oranı	14.5%
Yabancı bir firma tarafından kazanılan ihalelerin oranı	3.3%
Bir ihale için en az sayıda alınan teklif	1
Bir ihale için en çok sayıda alınan teklif	18
Ortalama teklif sayısı	3.918
Teklif sayılarının medyan değeri	3
Teklif sayılarının standart sapması	3.079

[3] https://data.europa.eu/euodp

3.1 Veri Kümesinin Oluşturulması

TED, sözleşme dokümanlarının XML formatında toplu halde indirilmesine imkan tanımaktadır. Veritabanında bulunan dokümanlar 22 ayrı sınıfta gruplandırılmıştır. Bu sınıflar Ek kısmındaki Çizelge 7.1'de görülebilir. Bu doküman çeşitlerinden sadece 3 kodlu "Contract Notice" başlıklı dokümanlar derlenmiştir. Veri kümesinin derlenme şeması Şekil 3.1'de görülebilir. Her ne kadar dokümanlar düzenli bir biçime sahip olsa da (ana başlıklar ve genel yapı gibi) yıllara göre XML doküman yapılarında farklılıklar bulunmaktadır. Ayrıca dokümanlardaki bilgi eksiklikleri ve hatalı bilgiler veri kümesinin oluşturmasını zorlaştırmaktadır. Örnek vermek gerekirse açıklama kısımlarını temsil eden tek bir XML etiketi bulunmamaktadır. Fakat dokümanların örnekler alınarak incelenmesiyle açıklamaların çoğunluğunun short_descr, short_description_contract ve short_contract_description olmak üzere üç etikette toplandığı görülmüştür. Diğer bir kısım dokümanlarda ise açıklamaların özel etiket isimleriyle ulaşılabildiği görülmüştür. Bu dokümanların da veri kümesine dahil edilmesi amacıyla İngilizce açıklama anlamına gelen "description" kelimesinin ve dokümanın orjinal yazıldığı dildeki "açıklama" anlamına gelen kelimelerin (örneğin, "Beskrivelse" Danimarka dilinde ve "Opis" Polonya dilinde) etiketlerde geçip geçmediği aranmıştır. Bu kelimeleri içeren etiketlerin alanlarındaki metinler de açıklama olarak alınıp veri kümesine dahil edilmiştir. Her bir dil için belirlenen bu kelimeler Ek kısmındaki Çizelge 7.2'de verilmiştir.



Şekil 3.1: Ham veriden kullanılan veri kümesinin oluşturulması gösteren akış şeması

Yayınlanan bazı dokümanlar birden çok dilde açıklama alanları bulundurmaktadır. Bunlardan bazıları orjinal dilde iken diğerleri orjinal metinlerin diğer dillerdeki tercümelerinden oluşmaktadır. Veri kümesi oluşturulurken orjinal dildeki açıklamalar kullanılmıştır. Eğer orjinal dil birden fazla dili içeriyorsa ve bu dillerden biri İngilizce ise İngilizce, değilse orjinal diller arasından rastgele bir tanesi seçilmiştir. Hazırladığımız veri kümesinde tercüme açıklama metinleri bulunmamaktadır. Ayrıca aynı ihaleye ait farklı dillerde metinler de bulunmamakta bu sayede içerik açısından aşırı örneklemeden (oversampling) kaçınmış olunmaktadır.

Yapılan ihalelerin farklı kısımları aynı ihale dokümanında farklı işler olarak belirtilebilir. Bu alt işler ihalelerde lot olarak tanımlanmaktadır. Farklı lotlara ait

kısımlar farklı firmalar tarafından kazanılabilir. Lotlara bölünmüş dokümanlarda ana açıklamalara ek olarak her bir lot için ana açıklamaya ek olarak lota özel açıklamalar bulunabilir. Bazı dokümanlarda ise ihale lotlara bölünmesine rağmen lot açıklamaları hiç olmayabilir veya lot açıklamaları aynı olabilir. Lot açıklamaları ana açıklama metinlerine ek olarak dokümanlardan elde edilmesine rağmen analizlerde kullanılmamıştır. Bu sayede açıklamaların muhtemel tekrarından dolayı aşırı örneklemeden kaçınmış olunmaktadır. Lot açıklamaların sözleşme sonuçlarına etkisi ileri bir araştırma konusu olarak (future work) incelenebilir.

Dokümanlardan elde edilen metinler doküman numaraları kullanılarak yayınlanan veri setleri ile birleştirilmiştir. Yayınlanan veri kümesindeki alanlardan ise sektör kodları, teklif sayısı, kazanan firmanın ülkesi, ihaleye çıkan otoritesinin ülkesi, tahmin edilen ihale fiyatı ve gerçekleşen ihale fiyatı bilgileri kullanılmıştır. Birleştirme işlemi sonrasında bu alanlardan herhangi birinde bilgi eksikliği bulunan satırlar düşürülmüştür. Fakat ihale tahmin fiyatı ve gerçekleşen fiyat alanlarındaki bilgi eksiklikleri diğer alanlardaki bilgi eksikliklerinden fazla olması sebebiyle teklif sayısı ve kazanan yabancı ülke analizlerindeki veri kümesini küçültmemek amacıyla iki ayrı veri kümesi oluşturulmuştur. Son durumda Tek Teklif, Teklif Sayısı ve Yabancı Firma problemleri için 333,832 satır, Fiyat Etkinliği problemi için 106,122 satır veri bulunmaktadır.

Teklif sayısı sütunundaki aykırı değerler de (outlier) ortalama değerden 5 standart sapma fazla olanlar çıkarılmak koşuluyla nihai veri kümesinde düşürülmüştür. Ayrıca teklif sayısında bilinmeyen değerler için kullanılan 999 değerine sahip olan satırlar da veri kümesinden çıkarılmıştır.

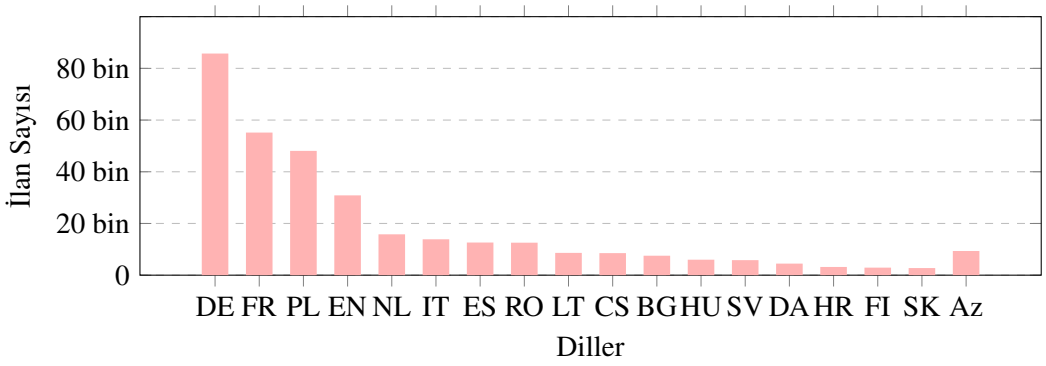
Hazırlanan veri kümesinin ön işleme aşamasında metinlerde XML formatında bulunan özel karakterler (&amp, &lt, &quot gibi) ham halden metne çevrilmiştir. Metinlerde sadece bu işlem uygulanmıştır. Nihai veri kümesinin farklı dillerdeki örnekleri Çizelge 3.2'de gösterilmiştir.

3.2 Diller

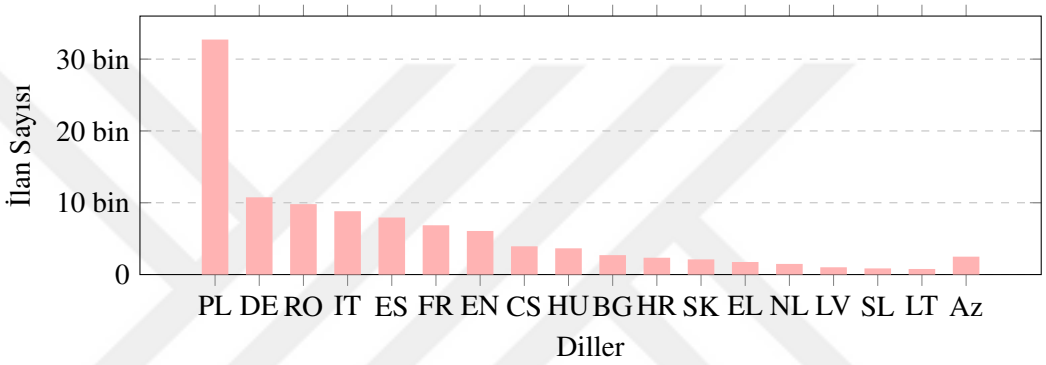
TED kamu alım ihale dokümanları 24 ayrı orjinal dilde bulunmaktadır. Bunlar diller Ekler kısmında Çizelge 7.2'de verilmiştir. Bu çalışma için ön işleme süreci sonunda oluşan veri kümesinde Maltaca ve İrlandaca hariç olmak üzere geri kalan 22 ayrı dilde açıklama metinleri bulunmaktadır. Bu iki dilde orjinal olarak yazılan metinlerin dokümanlarda oldukça kısıtlı olması sebebiyle az miktarda örneği veri kümesinden çıkarılmştır. Kalan dillerin dağılımı ise dengesiz bir özellik göstermektedir. Şöyle ki Almanca, Fransızca, Romence, Lehçe, İtalyanca ve İspanyolca açıklama metinleri veri kümesinin %71.72 sini oluşturmaktadır. Ayrıca 22 ayrı orijnal dili kapsamasıyla veri kümesi birbirinden oldukça farklı Latin, Kiril veya Helen gibi alfabelerle yazılmış metinlerden oluşmaktadır. Ayrıca birbirine yakın alfabeleri kullanan dillerde ise karakter farklılıkları bulunmaktadır. Veri kümesindeki dillere ait dağılımlar Şekil 3.2'de ve Şekil 3.3'te gösterilmiştir.

Çizelge 3.2: Lehçe, İngilizce ve Fransızca dillerinde örnek kamu ihale çağrıları.

İhale Açıklaması	**İhale Bilgileri**
Przedmiot zamówienia dotyczy dostawy kompleksowego wyposażenia banku komórek i zarodków wraz z adaptacją pomieszczeń w budynku IBD. W ramach dostawy wykonawca zobowiązany będzie do: wykonania dokumentacji projektowej, realizacji prac budowlanych, dostarczenia urządzeń do miejsca instalacji w siedzibie Zamawiającego i ich instalacji, do przeprowadzenia szkolenia pracowników Zamawiającego w zakresie obsługi urządzeń oraz uzyskanie pozwolenia na użytkowanie pomieszczenia.	**ID:** 2012197099 **CPV:** 42500000 **Otorite:** Instytut Biologii Doświadczalnej imienia Marcelego Nenckiego Polskiej Akademii Nauk (PL) **Kazanan Firma:** TK Biotech Tomasz Kamiński (PL) **Tahmin Fiyatı:** 174942.05 **Gerçekleşen Fiyat:** 159975.15 **Teklif Sayısı:** 2
The main objective of this study is to gather information on the dynamics between soil and water, focusing particularly on soil water retention. There is a need for detailed and systematic information on the effects that current changes in land use and increasing variations in precipitation and temperature have on this vital soil hydraulic property and on related components of the water cycle such as water run-off, percolation and infiltration. Single contract of 12 months with a total budget 92 000 EUR–115 000 EUR,	**ID:** 2012160673 **CPV:** 90700000 **Otorite:** European Commission, Directorates-General for the Environment/Climate Action, ENV.SRD.2 — Finance (BE) **Kazanan Firma:** Bio Intelligence Service (FR) **Tahmin Fiyatı:** 115000 **Gerçekleşen Fiyat:** 103025 **Teklif Sayısı:** 4
Souscription d'un contrat d'assurance groupe à adhésion facultative garantissant les risques statutaires des collectivités et établissements affiliés et non affiliés au Centre de Gestion vis-à-vis de leurs agents et des propres agents du Centre de Gestion (marché public à tranche ferme à bons de commande et tranches conditionnelles). Lot unique — «Risques statutaires du personnel» marché public à tranche ferme à bons de commande et tranches conditionnelles tranche ferme: collectivités et établissements de 20 agents et de moins de 20 agents Cnracl Tranches conditionnelles: collectivités et établissements de plus de 20 agents CNRACL	**ID:** 2016104952 **CPV:** 66512000 **Otorite:** Centre de gestion du Jura Fonction Pub (FR) **Kazanan Firma:** CNP Assurances (FR) — Sofaxis (FR) **Tahmin Fiyatı:** 6167578 **Gerçekleşen Fiyat:** 6167578 **Teklif Sayısı:** 1



Şekil 3.2: Tek Teklif, Teklif Sayısı ve Yabancı Firma problemleri için her bir dildeki kamu alım ilan sayılarının dağılımı. Az. veri kümesinde az miktarda olan son 5 dili göstermektedir. (Bu diller sırasıyla EL, LV, ET, SL, PT)



Şekil 3.3: Fiyat Etkin problemi için her bir dildeki kamu alım ilan sayılarının dağılımı. Az. veri kümesinde az miktarda olan son 5 dili göstermektedir. (Bu diller sırasıyla SV, DA, FI, PT, ET)

3.3 Sektör Kodu (CPV) (Common Procurement Vocabulary)

Sözleşme ilanları TED tarafından ait oldukları sektörlere göre hiyerarşik bir şekilde sırasıyla bölümler (divisions), gruplar (groups), sınıflar (classes), kategoriler (kategories), alt kategoriler (sub-categories) olmak üzere ayrılmıştır. 9 rakamdan oluşan bir CPV kodunun ilgili kısımları söz konusu ihalenin hangi sektöre ait olduğunu temsil etmektedir. İnşaat sektörü için örnek bir CPV kodu Çizelge 3.3'te gösterilmiştir.

Kullanılan veri kümesi toplamda 46 bölüm, 322 grup, 1304 sınıf, 3403 kategori ve 7841 alt-kategori içermektedir. İhalelerdeki söz konusu işler Berlin'deki silahlı kuvvetlere gıda teslimatından (ID: 2012385044) Vilnius'daki üniversite hastanesi için inşaat malzemeleri alımına kadar (ID: 201286594) geniş bir sektör dağılımı göstermektedir.

Çizelge 3.3: Örnek bir CPV kodunun ayrıştırılması. İlgili kısmı temsil eden rakamlar **kalın** olarak gösterilmiştir.

CPV Kısmı	**CPV Kodu**	**Ürün/Hizmet Tanımı**
Division	**45**000000-7	Construction work
Group	45**2**00000-9	Works for complete or part construction and civil engineering work
Class	452**2**0000-5	Engineering works and construction works
Category	4522**1**000-2	Construction work for bridges and tunnels, shafts and subways
Subcategory	45221**100**-3	Construction work for bridges
Subcategory	45221**110**-6	Bridge construction work
Subcategory	45221**111**-3	Road bridge construction work

4. TAHMİN PROBLEMLERİ VE ÖZELLİKLERİ

Bu çalışmanın ana motivasyonu kamu alım ilan metinlerinden ihale süreçlerinin AB komisyon direktiflerine uygunluğunu tespit edebilmektir. Fakat yüz binlerce ihalenin değerlendirilmesi, uygun olup olmadığının etiketlenmesi başlı başına yüksek kaynak ve zaman ayrılması gereken ve özellikle uzmanların yardımlarına ihtiyaç duyulan bir süreçtir. Bu nedenle doğrudan metinlerin uygunluğunu incelemek yerine ekonomistler tarafından kullanılan ekonomik ölçütler ihale süreçlerinin uygunluğunun ölçülmesinde kullanılmıştır. İhalelerin tek teklif alıp almadığı, toplam teklif sayısı, kazanan firmanın ihaleye çıkan otoritenin ülkesinden farklı bir ülkeden olması ve ihale fiyatının tahmin edilen fiyatı aşıp aşmadığı bilgisi olarak 4 adet ölçüt sadece metinler kullanılarak tahmin edilmektedir.

4.1 Teklif Sayısının Tahmini (Teklif Sayısı)

AB otoriteleri açık market kurallarına uygun olarak rekabetin korunması ve kamu kaynaklarının etkin bir şekilde kullanılmasını sağlamayı hedeflemektedirler. Teklif sayısının fazla olması ihale sürecindeki rekabeti artırmakla ihalesi için çıkılan ürün veya hizmetin daha uygun fiyatlarla otoriteler tarafından temin edilmesini sağlamaktadır. Bu nedenle teklif sayısının tahmin edilebilmesi ihale süreçlerinin rekabetçiliği ve fiyat etkin olmasında önemli bir ölçüttür [21]. Bu ölçüt sürekli (continuous) değişken olarak ele alınmış ve bu problem regresyon problemi olarak modellenmiştir. Çalışma boyunca bu problemden Teklif Sayısı ismiyle bahsedilecektir. Kullanılan veri kümesindeki teklif sayıları 3.079 standart sapma ile 3.918 ortalama değere sahip olup en az 1 en çok 18 teklif alan ihaleler veri kümesinde bulunmaktadır.

4.2 Tek Teklifli İhalelerin Tahmini (Tek Teklif)

İhaleler hakkında herhangi bir mahkeme kararı olmaksızın ihaleye fesat karıştırma hükmü vermek mümkün olmamasına rağmen tek teklifli ihaleler süreçlerin uygunsuz ilerlemesi konusunda şüphe uyandırmaktadır. Bu problemde, bir önceki probleme ek olarak teklif sayısı 1 veya 1'den fazla olmak üzere ikili (binary) bir değişken olarak incelenmiş ve problem ikili sınıflandırma (binary classification) problemi olarak modellenmiştir. Çalışma boyunca bu problemden Tek Teklif ismiyle bahsedilecektir. Kullanılan veri kümesindeki ihalelerin yüzde 21.0'lik kısmını tek teklif alan ihaleler oluşturmaktadır. Tek teklif belirlenmesinde kullanılan eşitlik Formül 4.1'de gösterilmiştir.

$$Hedef = \begin{cases} 1, & \text{Eğer teklif sayısı 1 ise} \\ 0, & \text{Diğer} \end{cases} \tag{4.1}$$

4.3 Yabancı Ülke Kaynaklı Tekliflerin Kazandığı İhaleler (Yabancı Firma)

Herhangi bir firma AB ülkelerinde olup olmaksızın, AB ülkelerinde düzenlenen ihalelere katılıp teklif verebilmektedir. Fakat doğal olarak ihale süreci yabancı firmalar için yerli firmalara kıyasla daha zorlayıcı olabilmektedir. Dil ve kültür farklılıkları ve hukuki gerekliliklere hakim olamama gibi nedenlerden dolayı yabancı firmalar kazanan firmalardan daha yetkin olmasına rağmen yabancı ülkelerdeki ihalelere daha az ilgiyle yaklaşabilirler. Diğer taraftan otoriteler bu durumu kötüye kullanarak ilanlarda daha az bilgi vermek suretiyle veya ürün ve hizmetle ilgili açıklamaları muğlaklaştırarak yabancı firmaların katılmasını engelleyebilirler. Buna binaen ekonomi literatüründe [17] yabancı firmalar tarafından kazanılan ihaleler rekabeti ölçen bir ölçüt olarak kullanılmaktadır. Bu problem de önceki problem gibi ikili sınıflandırma problemi şeklinde modellenmiştir. Çalışmanın sonraki kısımlarında bu problemden Yabancı Firma ismiyle bahsedilecektir. Kullanılan veri kümesindeki ihalelerin yüzde 3.3'lük kısmını yabancı firmalar tarafından kazanılan ihaleler oluşturmaktadır. Yabancı kazananın belirlenmesinde kullanılan eşitlik Formül 4.2'de gösterilmiştir.

$$Hedef = \begin{cases} 1, & \text{Eğer otorite ülkesi kazanan firmaların ülkeleri listesi içinde ise} \\ 0, & \text{Diğer} \end{cases} \tag{4.2}$$

4.4 Kazanılan Fiyatı Tahmin Edilen Fiyattan Fazla Olan İhalelerin Tahmini (Fiyat Etkin)

AB kamu alım otoriteleri ihaleler için talep edilen ürün veya hizmetlere karşılık gelen tahmini fiyat değerlerini açıklamaktadırlar. Japonya [14], İtalya [4], Türkiye [22], ve AB [9] piyasalarında yapılan araştırmalara göre genel olarak ihalelerin kazanım fiyatı otoriteler tarafından tahmin edilen fiyatın %90-95'ine denk gelmektedir. Ekonomi literatüründe tahmin edilen fiyattan yüksek fiyatta kazanılan ihaleler etkin olmayan ihale sürecine (ineffective bidding process) işaret etmektedir [30]. Bu problem de önceki sınıflandırma problemleri gibi ikili sınıflandırma olarak modellenmiş olup tahmin edilen fiyatla tamamen aynı kazanılan fiyata sahip ihaleler fiyat etkin sınıfına dahil edilmiştir. Çalışmanın sonraki kısımlarında bu problemden Fiyat Etkin ismiyle bahsedilecektir. Kullanılan veri kümesindeki ihalelerin yüzde 14.5'lük kısmını fiyat etkin olmayan ihaleler oluşturmaktadır. Fiyat etkinliğinin belirlenmesinde kullanılan eşitlik Formül 4.3'te gösterilmiştir.

$$Hedef = \begin{cases} 1, & \text{Eğer tahmin edilen fiyat gerçekleşen fiyattan büyük ise} \\ 0, & \text{Diğer} \end{cases} \tag{4.3}$$

5. ÖNERİLEN YÖNTEM

Bahsi geçen tahmin problemlerine genel olarak üç farklı yaklaşım geliştirilmiştir. İlk yaklaşım olarak çok dilli hassas ayarlanan dönüştürücü modeller incelenmiş, ikinci yaklaşım olarak çok dilli hassas ayarlanan dönüştürücü modelleri kullanılarak ihale metinlerinin cümle vektör gösterimleri elde edilmiş ve bu vektör gösterimleri ile makine öğrenmesi modelleri eğitilmiştir. Bu yaklaşıma ek olarak önemli bir gösterge olan CPV kodları kullanılarak aynı sektörde olan örnekler için ayrı modeller eğitilmiştir. Üçüncü yaklaşım olarak ihale metinlerinden adlandırılmış varlıklar, cümlenin ögeleri gibi dilbilgisel öznitelikler çıkarılmış ve bu öznitelikler kullanılarak tahminler yapılmıştır.

5.1 Çok Dilli Dönüştürücü Modellerin Hassas Ayarlanması Yaklaşımı

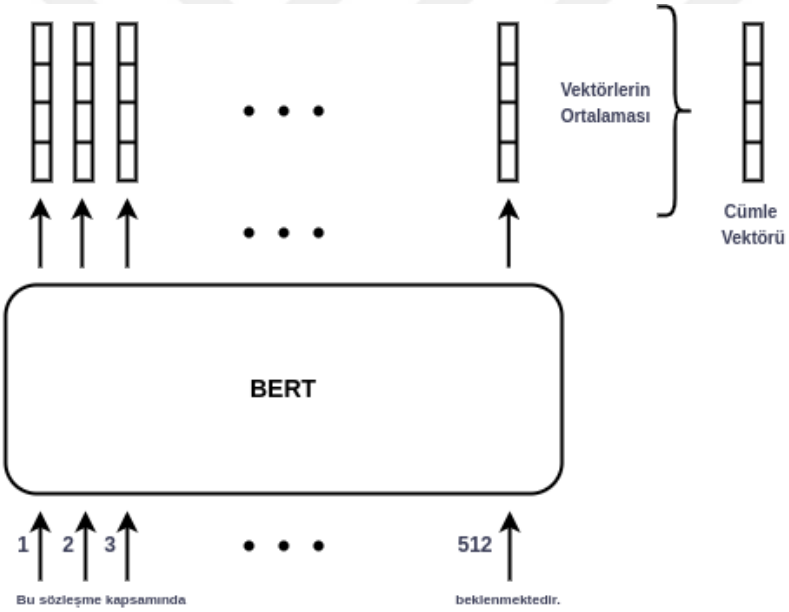
Deneylerde kullanılan veri kümesi 22 farklı Avrupa dilini kapsaması sebebiyle bütün metinleri girdi olarak alabilecek çok dilli dil modellerine ihtiyaç duymaktadır. Bu nedenle önceden çok dilli veri setlerinde eğitilmiş (pretrained) dönüştürücü dil modellerinin söz konusu problemler üzerinde hassas ayarlanması işlemi gerçekleştirilmiştir. Dönüştürücü dil modelleri olarak MBERT (Multilingual BERT) [7] ve XLMR [3] olmak üzere performanslarını karşılaştırabilmek amacıyla önceden eğitilmiş iki farklı mimariye sahip iki dönüştürücü dil modeli kullanılmıştır. Dönüştürücü modellerin girdi kapasiteleri göz önüne alındığında model girdi kapasitesini aşan metinler için kesme (truncation) ve model girdi kapasitesinin altında kalan örnekler için dolgu (padding) yöntemleri uygulanmıştır. Her bir problem için kullanılan modellerdeki tek farklı nokta son katmanda (layer) regresyon problemi için lineer katman, sınıflandırma problemleri için ise lineer katmana ek olarak sigmoid işlemi gerçekleştiren sınıflandırma katmanı eklenmiş olmasıdır. Kayıp (loss) fonksiyonu olarak ise regresyon problemi için MAE (mean squared error), sınıflandırma problemi için ise CE (cross-entropy) kullanılmıştır.

5.2 Cümle Vektör Gösterimlerinin Kullanılması Yaklaşımı

Önceden eğitilmiş dönüştürücü modelleri kullanılarak farklı dillerdeki ihale metinlerinin tek bir uzayda (latent space) vektör gösterimleri elde edilebilir. Bu sayede elde edilen vektör gösterimleri öznitelik olarak kullanılarak bahsi geçen problemler için modeller geliştirilebilir. Hassas ayarlanan çok dilli modeller için kullanılan MBERT ve XLMR modelleri vektör gösterimleri elde etmek için kullanılmıştır. Elde edilen vektör gösterimleri farklı makine öğrenmesi modelleri kullanılarak performansları karşılaştırılmıştır. Cümle vektör gösterimlerinin elde edilme şeması Şekil 5.1'de gösterilmiştir.

5.2.1 Sektör kodlarına göre filtrelenmesi

Veri kümesi kısmında bahsedildiği gibi, kullanılan veri kümesi oldukça geniş alanda yayılan ürün ve hizmetlerle alakalı dokümanlardan derlenmiştir. Bu nedenle ihale sonuçlarını etkileyecek en büyük etkenlerden biri talep edilen ürün veya hizmetlerin çeşidi olmaktadır. Örneğin, inşaat malzemelerinin satın alınmasıyla alakalı bir ihale için alınan teklif sayısı gıda malzemelerinin satın alınmasıyla alakalı bir ihale için alınan teklif sayısından daha düşük olacaktır. Eğer bir otorite talep ettiği özel bir malzemeyi (özel bir medikal cihaz gibi) kendi ülkesindeki firmalardan tedarik edemezse yabancı bir firmadan bu tedariki yapmak durumunda kalacaktır. Veya farklı sektörlerdeki ürün ve hizmetlerdeki kar marjları her sektör için aynı olmayacaktır. Böylece daha uygun teklif verebilecek firmalar farklı sektörlerde farklı miktarlarda olacak ve otoriteler talep ettikleri ürün ve hizmetleri, o ürün ve hizmetlerin çeşidine göre farklı fiyat etkinlik oranlarıyla satın alabileceklerdir. Bu şekildeki makroekonomik etkileri de göz önünde bulundurmak amacıyla eğitim kümesi her bir sektör bölümüne göre (division, CPV kodunun ilk iki rakamıyla temsil edilmektedir) ayrılmıştır. Ayrılan her bir bölüm için modeller ayrıca eğitilmekte ve yine aynı sektörler için olan örneklerde tahmin sonuçları alınmaktadır.



Şekil 5.1: MBERT ve XLMR modelleri kullanılarak cümle vektör gösterimlerinin oluşturulma şeması

5.3 Öznitelik Tabanlı Yaklaşım

Bu yaklaşımda ihale metinlerinin anlamlı bir uzayda belirli özniteliklerle temsil edilip öğrenme modelleri geliştirilmesi benimsenmiştir. Öznitelik mühendisliği aşamasında elde edilen öznitelikler başlıca dilbilgisel ve dilbilgisel olmayan öznitelikler olarak iki kısma ayrılabilir. Dilbilgisel olarak ihale metninin yazıldığı dil, ihale metninin karakter ve kelime bazlı uzunluğu, adlandırılmış varlıklar, cümlenin ögeleri, kelime ve kelime öbeklerinin frekansları öznitelikleri kullanılmıştır. Dilbilgisel olmayan öznitelikleri ise otoritenin ülkesi, talep edilen ürün veya hizmetin dahil olduğu sektör ve alt sektörler

olarak sıralanabilir.

Deneylerde kullanılan veri kümesinin 22 farklı dilde metinlerden oluşması sebebiyle dilbilgisel özniteliklerin çıkarılması her bir dil için farklı dil modelleri gerektirmektedir. Fakat birçok doğal dil işleme kütüphanesi İngilizce'ye ve popüler olan bir kaç dile odaklanmakta olup diğer dillerde doğal dil işleme araçları bulmak mümkün olmamaktadır. Bu nedenle uygun dil araçlarının bulunmadığı diller için metinler Google Translate[4] kullanılarak İngilizce'ye tercüme edilmiş ve İngilizce doğal dil işleme araçları kullanılmıştır.

5.3.1 Dilbilgisel öznitelikler

- **İhale metninin yazıldığı dil (LG):** İhalenin yazıldığı dil ihale sürecini büyük ölçüde etkileyebilir. Çünkü teklif veren yabancı firmalar ihale açıklamalarını tam olarak anlayamazlarsa ihaleden çekilebilirler. Ayrıca farklı ülkelerdeki farklı dillerin kullanım oranları o ülkenin kültürüne ve yabancı dillerin popülerliğine göre değişkenlik gösterebilmektedir. Örneğin çoğu insan ikinci bir dil olarak Macarca veya Litvanca yerine İngilizceyi öğrenmeyi tercih edecektir. Bazı ülkeler içinse dil farklılıkları herhangi bir problem arz etmeyecektir. Örnegin Almanya'da Almanca olarak yayınlanan bir ihale metni için Avusturyalı bir firma diğer ülkelerdeki rakipleri kadar zorluk yaşamayacaktır. Veri kümesinde Çizelge 7.2'de gösterilen 22 adet dil bulunmaktadır.
- **Açıklama metninin uzunluğu (LEN):** Açıklama metninin detaylı hazırlanması katılımcı firmaların yükleneceği riskleri önceden tahmin edebilmesi ve belirsizliklerin ortadan kalmasıyla yapılacak işin tam olarak tanımlanması anlamına gelmektedir. Açıklama metinlerinin uzun olması daha çok detay içermesi hakkında fikir verebilir. Bu öznitelik karakter ve kelime bazlı olarak metinlerin uzunluklarını tutmaktadır.
- **Adlandırılmış varlıklar (NER):** Her ne kadar uzun metinlerin daha fazla bilgi içermesi beklense de göreceli olarak kısa metinlerin de uzun metinler kadar önemli bilgiler içermesi mümkündür. Metinlerde geçen kişi, mekan, kurum isimleri, tarih ve zaman bilgileri ihale hakkında önemli detaylar sunmaktadır. Bu öznitelik metinlerde geçen adlandırılmış varlıkları kişi (PER), yer (LOC), organizasyon (ORG) ve diğer (MISC) başlıkları altında 4 ana grupta toplamaktadır. "Diğer" başlığının altında sayı, tarih, para, yüzde, ürün, miktar, zaman gibi öznitelikler dahildir. Adlandırılmış varlıkların kısa açıklamaları Ek kısmında Çizelge 7.5'te ve her bir dil için etiketler Çizelge 7.6'da verilmiştir.
- **Cümlenin ögeleri (POS):** Adlandırılmış varlıklara benzer olarak cümlede geçen öge sayıları da metinlerin ne kadar bilgi içerdiği hakkında fikir verebilir. Örneğin, bir metin çok fazla isim geçmesi o metin hakkında daha çok bilgi içermesi anlamına gelebilir. Özne, yüklem, isim, bağlaç, zarf gibi 19 adet cümle ögeleri bulunma miktarlarına göre açıklama metnini temsil etmektedir. Cümlenin ögeleri için etiketler Ek kısmında Çizelge 7.4'de verilmiştir.
- **Kelimeler ve kelime öbekleri (NG):** Geçmişte olan benzer ihalelerin gelecekte de benzer sonuçlar vermesini öngörmekteyiz. Bu nedenle benzer işleri

[4] translate.google.com/

tanımlayan kelime ve kelime öbekleri ihale sonuçlarını tahmin etmek açısından önem taşımaktadır. Örneğin Berlin'deki 20km yol bakım çalışması, yine Berlin'deki 35km yol bakım çalışması ile benzer ihale sonuçları taşıdığı düşünülebilir. Kelime ve kelime öbekleri (n-grams) açıklama metinlerin bulunup bulunmama durumlarına göre o metni temsil edebilirler. Kelime teklileri ile beraber kelime öbekleri ikili kelime çiftleri halinde seçilmiş ve metinlerde geçme frekansları TF-IDF skorları üzerinden hesaplanmıştır.

5.3.2 Dilbilgisel olmayan öznitelikler

- **Sektör kodu (CPV):** Veri kümesi kısmında açıklandığı üzere sektör kodları TED tarafından ihaleler hakkında temin edilen en önemli bilgilerden biridir. Aynı sektördeki ihalelerin sonuçları açısından birbirine benzer olması beklenebilir. Örneğin, gıda sektöründe tedarik için çok fazla teklif alınırken, medikal sektöründe kazanan yabancı firma sayısı daha fazla olabilir. Her bir CPV kodunun her bir bölümü genelden özele olacak şekilde ihale metninin konusu hakkında bilgi vermektedir. Fakat alt kategorilere kadar bütün CPV kısımlarının kullanılması oldukça kısıtlayıcı olabilir. Bu nedenle daha genel bir gruplama yöntemi olarak sadece CPV bölümleri (ilk iki rakam) sektör kodu özniteliği olarak kullanılmıştır.
- **Otoritenin ülkesi (ISO):** Kamu alım otoriteler kendi ülkesinde firmalara daha toleranslı olabilirler. Örneğin Bulgaristan'daki bir altyapı ihalesi için Bulgar firmalar doğal olarak daha çok tercih sebebi olabilir. Çünkü insanları, coğrafyayı ve lojistik imkanları daha iyi kullanabilirler. Bu nedenle gelişmiş ülkelerdeki ihale katılım oranları ve kazanan firmanın ülkesi diğer daha az gelişmiş ülkelerden daha farklı olabilir. Bu öznitelik TED veri kümesinde belirtilen ülke ISO kodları ile temsil edilmektedir. Veri kümesinde 32 adet ülkeye ait otoritelere ait örnekler bulunmaktadır. Bu ülkelerin listesi Ek kısmında Çizelge 7.3'te gösterilmiştir.

6. DENEYLER

Tez çalışmasının bu bölümü çalışmada yapılan deneylerden bahsetmektedir. Deneyler sırasında kullanılan kütüphanelerden, oluşturulan temel modellerden, önerilen yaklaşımların karşılaştırılmasından ve deney sonuçlarının yorumlanmasından bahsetmektedir.

6.1 Deney Kurulumu

Veri kümesinin oluşturulmasıyla ilgili **Bölüm 3**'te bahsedilmiştir. Bu kısımda ise veri kümesinin bölümlenmesi, modelleri karşılaştırırken belirlenen ölçütler ve model uygulama detaylarından bahsedilecektir.

6.1.1 Eğitim, değerlendirme ve test kümeleri

Veri kümesi kısmında bahsedildiği gibi deneylerde iki farklı veri kümesi kullanılmıştır. Teklif Sayısı, Tek Teklif, Yabancı Firma problemleri için kullanılan veri kümesi 333.832 örnek içerirken Fiyat Etkin problemi için kullaılan veri kümesi 106.122 örnek içermektedir. Her bir veri kümesi için %80'lik kısım eğitim kümesine ayrılırken kalan kısım test için ayrılmıştır. Eğitim ve test kümelerine ayrılmadan önce veri kümesi kendi içerisinde karma işlemi (shuffling) gerçekleştirilmiştir. Parametre optimizasyonları için ise eğitim kümesinin %75'lik ilk kısmı eğitim %25'lik son kısmı değerlendirme kümesi olarak kullanılmıştır.

6.1.2 Değerlendirme ölçütleri

Modellerin başarımları değerlendirilirken sınıflandırma modelleri için F_1 skoru, regresyon modelleri için MAE skoru kullanılmıştır. F_1 skoru hesaplanırken Tek Teklif problemi için tek teklifli ihaleler, Yabancı Firma problemi için yabancı firmanın kazandığı ihaleler, Fiyat Etkin problemi için ise fiyat etkin olmayan ihaleler yani azınlık olan sınıflar pozitif sınıflar olarak belirlenmiştir.

6.1.3 Temel modeller

Önerilen modellerin performanslarını karşılaştırmak amacıyla basit medyan gruplama modellerinden metinleri ve ekonomik öznitelikleri kullanan modellere çeşitli yaklaşımlar kullanılmıştır.

- **Eğitim Kümesinin Medyanı** ($Medyan_{Egitim}$): Eğitim kümesindeki bütün örneklerin medyanını almak en temel yöntem olarak hesaplanmıştır. Fakat bu temel model sadece Teklif Sayısı problemi için değerlendirilecektir.

- **Otorite Ülkesi Aynı Olan İhalelerin Medyanı** ($Medyan_{ISO}$: Aynı ülke otoriteleri tarafından düzenlenen ihaleler için medyan değerini almak ikinci temel yöntem olarak hesaplanmıştır. Fakat bu yöntem de büyük ölçüde Teklif Sayısı problemi için değerlendirmeye alınacaktır.

- **Sektörü Aynı Olan İhalelerin Medyanı** ($Medyan_{CPV}$): Sektör bilgisinin ihale sonuçları açısından etkilidir. Benzer sektördeki ihaleler benzer sonuçlar verebilir. Bu temel modelde alt kategori CPV seviyesi kullanılmıştır. Bu sayede sektör açısından birbirine en çok benzeyen ihalelerin medyan değerleriyle tahmin yapılmaktadır. Eğitim kümesinde bulunmayan sektörler için CPV bölüm seviyesi alınmaktadır.

- **Otorite Ülkesi ve Sektörü Aynı Olan İhalelerin Medyanı** ($Medyan_{CPV\_ISO}$): Aynı ülke otoritesine ve aynı sektöre sahip ihalelerin medyanının alınmasıyla tahmin yapılmaktadır. Önceki iki temel modelin birleşimi olarak düşünülebilir.

- **Tek kelime öznitelikleriyle yakın komşuluk modeli** ($KNN_{TekKelime}$): Görgün vd. çalışmasında [12] KNN modelinin kelime öznitelikleriyle TED veri kümesi üzerinde teklif sayısını tahmin etmede etkili olduğu gösterilmiştir. Bu çalışmada ise daha geniş bir veri kümesi ile 10.000 adet tek kelime (uni-gram) öznitelikleri kullanılarak temel model olarak KNN modeli çalıştırılmıştır.

- **Tek kelime öznitelikleriyle Lineer modelleme** ($LR_{TekKelime}$): Bir önceki temel modele benzer şekilde aynı özniteliklerin lineer modellemesi bu temel modelde gerçekleştirilmiştir.

- **Makroekonomik özniteliklerle Lineer modelleme** ($LR_{Makroekonomik}$): Teklif sayısını etkileyen etmenleri incelemek üzere Tas vd.[29] çalışmasında ihaleye çıkan otoritenin tipi, ülkesi, CPV bölümü (ilk iki rakam), prosedür tipi ve sözleşmenin tedarik anlaşması güvencesinde yapılıp yapılmadığı değişkenleri genelleştirilmiş lineer modelleme yöntemiyle modellenmiştir. Bu çalışmada kullanılan değişkenleri öznitelikler olarak kullanarak lineer modelleme yöntemiyle kendi problemlerimizi tahmin için temel model oluşturulmuştr..

6.1.4 Uygulama detayları

Kelime öznitelikleri oluşturulurken Scikit-Learn kütüphanesinin[5] öznitelik çıkarımı modülünün TFIDFVectorizer aracı kullanılmıştır. Vektorizasyon işlemi öncesinde token ayırma işlemi için NLTK[6] kütüphanesinin RegexpTokenizer modülü kullanılmıştır. Minimum doküman frekans parametresi ise 0,8 olarak belirlenmiştir. Temel modellerde kullanılan parametreler için kütüphanelerdeki varsayılan değerler kullanılmıştır. Öznitelik sayısı belirlenirken ise 5000, 10000 ve 20000 değerleri tek kelime ve iki kelime öznitelikleri için karşılaştırılmış ve temel modellerde en iyi sonucu veren

[5]https://scikit-learn.org/stable/index.html
[6]https://www.nltk.org/

10.000 adet tek kelime öznitelikleri kullanılmıştır. Parametre seçimi için yapılan karşılaştırma sonuçları Çizelge 6.1'de gösterilmiştir. Karşılaştırmalar Teklif Sayısı problemi için lineer modelle yapılmıştır.

Çizelge 6.1: Kelime ve kelime öbekleri öznitelikleri için vektör uzunluğu seçimi sonuçları

	5000	10000	20000
bi-gram	2.048	2.036	2.056
uni-gram	1.970	**1.955**	1.970

Çok dilli dönüştürücü modellerin kullanıldığı yaklaşımlarda Huggingface Transformers[7] kütüphanesinden faydalanılmıştır. Cümle vektör gösterimlerinin kullanımı yaklaşımında MBERT ve XLMR modellerinden elde edilen vektör gösterimleri kullanılarak k en yakın komşu (KNN_{MBERT}, KNN_{XLMR}), lineer regresyon (LR_{MBERT}, LR_{XLMR}), rastgele orman (RF_{MBERT}, RF_{XLMR}) ve çok katmanlı sinir ağı (MLP_{MBERT}, MLP_{XLMR}) modelleri eğitilmiş ve her bir tahmin problemi için performansları ölçülmüştür. Ayrıca aynı modellerin sektör filtreleme yöntemiyle performansları da ölçülüp karşılaştırılmaktadır.

Hiper parametre araması için değerlendirme kümesi olarak eğitim setinin 1/4'ü kullanılmakta olup test kümesi boyutuyla eşittir. Hassas ayarlanan modellerde her bir problem için 3 epoch ve 32 batch boyutu sabit olmak üzere 1e-5, 2e,5 ve 5e-5 değerleri için öğrenme oranı ayarlanmıştır. Değerlendirme kümesindeki en iyi değerler Çizelge 6.2'de gösterilmiştir. Buna göre öğrenme oranı Tek Teklif, Teklif Sayısı, Yabancı Firma problemleri için 1e-5, Fiyat Etkin problemi için 2e-5 olarak ayarlanmıştır. Öğrenme oranı ayarlaması sadece MBERT ile yapılmıştır.

Çizelge 6.2: Öğrenme oranının değerlendirme kümesinde ayarlanması: En iyi değerler **kalın** ile gösterilmiştir.

	Tek Teklif	Teklif Sayısı	Yabancı Firma	Fiyat Etkin
1e-5	**0.436**	**1.827**	**0.399**	0.189
2e-5	0.432	1.838	0.381	**0.198**
5e-5	0.000	1.870	0.000	0.151

6.2 Deney Sonuçları

Bu bölümde temel model performansları, önerilen yöntemlerin temel modeller ile karşılaştırılması, sektör filtrelemesinin model performansları üzerine etkisi, çok dilli eğitimin etkisi ve dil geçişli eğitimin etkisi raporlanmaktadır.

6.2.1 Temel model performansları

Kullanılan temel modellerin performansları Çizelge 6.3'te gösterilmiştir. $Medyan_{Egitim}$ modeli ikili sınıflandırma problemleri için beklenen şekilde 0 F_1 skoruna sahiptir. Ülke

[7] https://huggingface.co/models

bazlı gruplama yapılan $Medyan_{ISO}$ temel modeli ise ikili sınıflandırma problemlerinde Fiyat Etkin için 0'dan farklı bir F_1 skoruna sahiptir. Bu durum göstermektedir ki ülke bazında gruplama yapıldığında eğitim kümesinin genelinde fiyat etkin ihaleler çoğunlukta olmasına rağmen fiyat etkin olmayan ihalelerin çoğunlukta olduğu ülkeler bulunmaktadır.

Çizelge 6.3: Temel modellerin karşılaştırmalı sonuçları. En iyi modeller **kalın** ile gösterilmiştir.

Metot	Tek Teklif	Teklif Sayısı	Yabancı Firma	Fiyat Etkin
$Medyan_{Egitim}$	0.000	2.160	0.000	0.000
$Medyan_{ISO}$	0.000	2.059	0.000	0.137
$Medyan_{CPV}$	0.235	1.912	0.042	0.036
$Medyan_{CPV\_ISO}$	**0.378**	**1.846**	**0.282**	0.141
$KNN_{TekKelime}$	0.366	1.975	0.223	**0.224**
$LR_{TekKelime}$	0.308	1.950	0.205	0.125
$LR_{Makroekonomik}$	0.246	2.037	0.190	0.119

Sektör bazlı gruplama ($Medyan_{CPV}$) yapıldığı zaman ise sonuçlarda 0 değerine rastlanmamaktadır. Bu ise farklı sektörlerde en çok geçen etiketin farklılaştığını göstermektedir. $Medyan_{CPV\_ISO}$ modeli ise dört problemden üçünde temel modeller arasında en iyi performansı göstermektedir. Ayrıca Teklif Sayısı problemi için Hassas ayarlanmış XLMR modelini ve MLP modellerini, Yabancı Firma problemini için ise XLMR vektörlerini kullanan MLP modellerini geride bırakmaktadır. Bu durum göstermektedir ki aynı ülke ve aynı sektörde yayınlanan ihaleler teklif sayısı ve yabancı firmaların kazanma durumu açısından oldukça benzer özellik göstermektedirler. Fiyat Etkin probleminde ise $KNN_{TekKelime}$ temel modeli metin içeriklerini kullanarak $Medyan_{CPV\_ISO}$ modelinden daha iyi performans göstermiştir. Kelime özniteliklerinin kullanıldığı iki tip modelden $KNN_{TekKelime}$ sınıflandırma problemlerinde daha iyi iken $LR_{TekKelime}$ regresyon probleminde daha iyi sonuç vermektedir. Yalnız makroekonomik özniteliklerin kullanıldığı lineer model olan $LR_{Makroekonomik}$ ise kelime öznitelikleri kullanan modellerin ve $Medyan_{CPV\_ISO}$ modelinin gerisinde kalmaktadır.

6.2.2 Önerilen yöntem performansları

Önerilen yöntem performansları ve temel model kıyaslamaları Çizelge 6.7'de verilmiştir. Her ne kadar temel modellerde $Medyan_{CPV\_ISO}$ temel modeli, metin kaynaklı kelime özniteliklerini kullanan $KNN_{TekKelime}$ ve $LR_{TekKelime}$ modellerinden daha iyi performans gösterse de önerilen yöntemler bütün problemlerde temel model performanslarını geride bırakmaktadır. Bu durum göstermektedir ki metin içerikleri ihale sonuçları üzerinde etkili olmaktadır.

Önerilen yöntemler arasında performansları açısından en başarılı olanlar çok dilli dönüştürücü model yaklaşımları kullanan yöntemler olmuştur. MBERT ve XLMR modelleri diğer yöntemlerle kıyaslanırken 3 farklı çalıştırma performansının ortalaması baz alınmıştır. Bu çalıştırma sonuçları Çizelge 6.4'te ve Çizelge 6.5'de görülebilir. MBERT modeli her bir problem için XLMR modelinden daha düşük varyans göstermektedir. Hassas ayarlama performansında ise MBERT modeli dört problemin

hepsinde XLMR'dan daha iyi sonuçlar vermektedir.

Çizelge 6.4: MBERT modeli için 3 farklı çalıştırma sonuçları

	Tek Teklif	Teklif Sayısı	Yabancı Firma	Fiyat Etkin
Tur 1	0.436	1.811	0.393	0.182
Tur 2	0.436	1.815	0.397	0.172
Tur 3	0.435	1.810	0.393	0.171
Ortalama	0.436	1.812	0.394	0.175
Standart Sapma	0.001	0.003	0.002	0.006

Çizelge 6.5: XLMR modeli için 3 farklı çalıştırma sonuçları

	Tek Teklif	Teklif Sayısı	Yabancı Firma	Fiyat Etkin
Tur 1	0.406	1.846	0.382	0.134
Tur 2	0.410	1.883	0.378	0.136
Tur 3	0.381	1.903	0.377	0.127
Ortalama	0.399	1.877	0.379	0.132
Standart Sapma	0.016	0.029	0.003	0.005

Öznitelik bazlı modeller ise önerilen yöntemler arasında genel olarak düşük performanslara sahiptir. Farklı öznitelikler gruplarının kıyaslaması Çizelge 6.6'da görülebilir. Dil özniteliğinin Tek Teklif ve Fiyat Etkin problemlerindeki başarısı temel yöntemlerdeki $KNN_{TekKelime}$ modeline paralellik göstermektedir. Bunun nedeni çoklu dilde vektörizasyon yaparken farklı dilde bulunan vektörlerin 0 yakınlık göstermesidir. Teklif Sayısı ve Yabancı Firma problemlerinde ise benzer şekilde ISO özniteliği ile $Medyan_{CPV\_ISO}$ paralellik göstermektedir.

Teklif Sayısı probleminde hassas ayarlanmış MBERT modeli bütün temel model performanslarını geride bırakmıştır. Cümle vektör gösterimlerinin kullanıldığı modeller ise en iyi temel model performansı gösteren $Medyan_{CPV\_ISO}$ skorunun gerisinde kalmaktadırlar. Cümle vektör gösterimlerinin kullanıldığı modellerde genel olarak XLMR vektör gösterimlerinin kullanıldığı modeller daha başarılı olduğu gözlemlenmektedir. Çizelge 6.8'de görülebileceği gibi sektör filtreleme ise genellikle performansı iyileştirici yönde etkilemiş fakat yalnızca MLP_{XLMR_CPV} modeli kullanıldığında $Medyan_{CPV}$ temel model performansı geçilebilmiştir.

Tek Teklif probleminde ise hassas ayarlanmış iki model ve sektör filtresi kullanılan iki

Çizelge 6.6: KNN modeli için öznitelik tabanlı model yaklaşımı sonuçları

Öznitelikler	Tek Teklif	Teklif Sayısı	Yabancı Firma	Fiyat Etkin
NER	0.185	2.602	0.001	0.048
POS	0.227	2.297	0.034	0.114
LEN	0.119	2.494	0.003	0.061
LG	**0.325**	2.750	0.000	**0.138**
CPV	0.089	2.425	0.000	0.027
ISO	0.059	**2.213**	**0.190**	**0.138**

Çizelge 6.7: Temel modeller ve önerilen yöntemlerin deney sonuçları. En iyi sonuçlar **kalın** ile gösterilmiştir.

Metot	Tek Teklif	Teklif Sayısı	Yabancı Firma	Fiyat Etkin
$Medyan_{CPV}$	0.235	1.912	0.042	0.036
$Medyan_{CPV\_ISO}$	0.378	1.846	0.282	0.141
$KNN_{TekKelime}$	0.366	1.975	0.223	0.224
$LR_{TekKelime}$	0.308	1.950	0.205	0.125
$LR_{Makroekonomik}$	0.246	2.037	0.190	0.119
MBERT	**0.436**	**1.812**	**0.394**	0.175
XLMR	0.399	1.877	0.379	0.132
MLP_{MBERT}	0.412	2.040	0.285	**0.286**
MLP_{MBERT_CPV}	0.430	1.998	0.305	0.236
MLP_{XLMR}	0.374	1.915	0.274	0.131
MLP_{XLMR_CPV}	0.397	1.853	0.261	0.124
$KNN_{Oznitelik}$	0.227	2.315	0.017	0.104
$GB_{Oznitelik}$	0.170	2.133	0.247	0.116

MLP modeli tüm temel modelleri geride bırakmıştır. Teklif Sayısı probleminden farklı olarak cümle vektör gösterimlerinin kullanıldığı KNN ve MLP modelleri bu problemde en iyi temel model performansını gösteren $Medyan_{CPV\_ISO}$'i geride bırakmaktadırlar. KNN modelinde XLMR cümle gösterimleri daha iyi performans sergilerken MLP modelinde MBERT vektörleri daha iyi sonuç vermektedir. Sektör filtrelemenin etkisi incelendiğinde ise bütün modellerde filtrelemenin önemli bir performans artışı sağladığı görülmektedir. Dikkat çeken diğer bir nokta ise LR_{MBERT} ve LR_{XLMR} modelleri $Medyan_{CPV}$ skorunun gerisinde kalırken filtreleme yapılan çeşitleri olan LR_{MBERT_CPV} ve LR_{XLMR_CPV} modelleri $Medyan_{CPV}$ skorunu geçmektedirler.

Yabancı Firma probleminde ise hassas ayarlanmış iki model ve MBERT cümle vektör gösterimlerinin kullanan KNN ve MLP modelleri tüm temel modelleri geride bırakmıştır. Hem KNN hem MLP modellerinde MBERT vektörleri XLMR vektörlerine göre daha iyi performans göstermektedir. Tek Teklif probleminde farklı olarak bu problemde sektör filtrelemesi bazı model ve vektör tipleri için performansı ters yönde etkilemekte olduğu gözlenmiştir. Örneğin MLP modelinde MBERT vektörleriyle filtreleme daha iyi sonuç verirken XLMR vektörleriyle filtreme ters yönde etkilemiştir. Fakat her iki KNN modelinde sektör filtreleme performansı azaltıcı yönde etkilemiştir. Bakıldığında bu problem için sektörün etkisinin az olması $Medyan_{CPV\_ISO}$ ve $Medyan_{CPV}$ farkından da anlaşılabilir. Modeller de bunu destekler yönde sonuçlar vermiştir.

Fiyat Etkin probleminde diğer problemlerden farklı şekilde en iyi performans gösteren modeller cümle gösterimlerini kullanan modeller olmuştur. Hassas ayarlanmış iki model de $KNN_{TekKelime}$ temel model performansının gerisinde kalmıştır. Cümle vektör gösterimlerinin kullanan modellerden MLP modeli MBERT vektörleri ile, KNN modeli ise XLMR vektörleriyle daha iyi sonuçlar vermektedir. Filtrelemenin etkisi incelendiğinde ise Yabancı Firma problemine benzer şekilde performansa etikisi ters yönlü olmuştur. $Medyan_{CPV\_ISO}$ ve $Medyan_{CPV}$ skor farkına bakıldığında bunun da beklenen bir sonuç olduğu söylenebilir.

Çizelge 6.8: Cümle vektör gösterimleri yaklaşımı sonuçları. En iyi sonuçlar **kalın** ile gösterilmiştir.

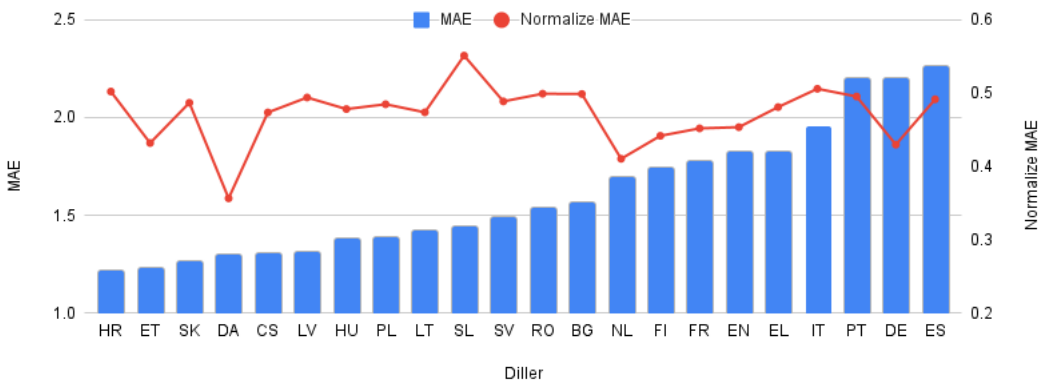
Method	Filtre	Tek Teklif	Teklif Sayısı	Yabancı Firma	Fiyat Etkin
KNN_{MBERT}	✗	0.397	2.127	0.300	0.268
	✓	0.412	2.092	0.293	0.267
KNN_{XLMR}	✗	0.400	2.129	0.294	0.277
	✓	**0.417**	2.085	0.288	0.281
LR_{MBERT}	✗	0.215	2.024	0.135	0.109
	✓	0.303	2.070	0.115	0.045
LR_{XLMR}	✗	0.058	1.999	0.000	0.000
	✓	0.143	2.046	0.000	0.000
RF_{MBERT}	✗	0.215	2.110	0.059	0.105
	✓	0.341	1.997	0.114	0.097
RF_{XLMR}	✗	0.260	2.082	0.116	0.123
	✓	0.361	1.981	0.142	0.124
MLP_{MBERT}	✗	0.412	2.040	0.285	**0.286**
	✓	0.430	1.998	**0.305**	0.236
MLP_{XLMR}	✗	0.374	1.915	0.274	0.131
	✓	0.397	**1.853**	0.261	0.124

6.2.3 Test setinde farklı dillerdeki model performansları

Kullanılan veri kümesinin çoklu dilde olmasının etkilerini gözlemlemek amacıyla hassas ayarlanmış MBERT modelinin test setinde bulunan farklı dillerdeki örnekler üzerindeki performansı ayrıca incelenmiştir.

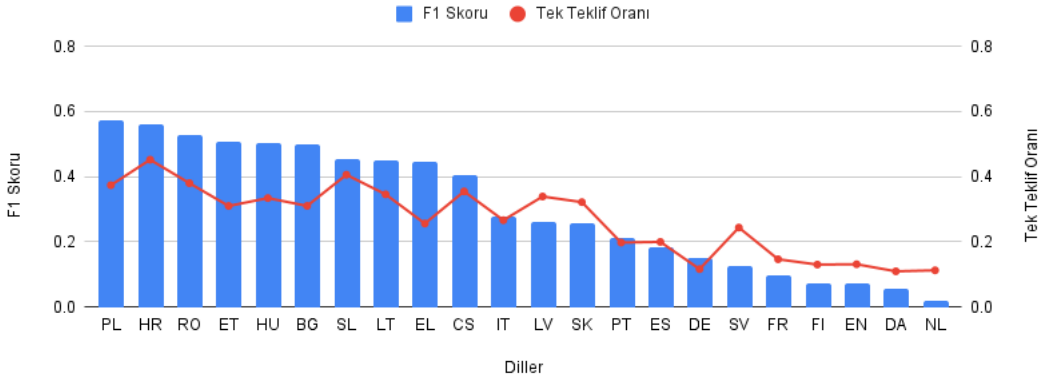
Teklif Sayısı problemindeki farklı dillerdeki model performansları Şekil 6.1'de gösterilmiştir. En düşük hata skorları Hırvatça, Estonca, Slovakça, Danca ve Çekçe'de gözlenirken, en yüksek hata skorları İspanyolca, Almanca, Portekizce, İtalyanca ve Yunanca'da gözlemlenmiştir. Genel olarak ekonomik olarak gelişmiş ülke dillerinde hata oranları yüksek iken ekonomik olarak görece daha az gelişmiş ülke dillerinde hata oranları daha düşüktür. Bunun sebebi ekonomik olarak gelişmiş ülkelerde ihale koşullarını sağlayabilen daha çok firma olmasıyla teklif sayısının artması olarak gösterilebilir. Teklif sayısındaki varyansın yüksek olması model performanslarını düşürmekte olabilir. Bu etkiden arındırmak için Şekil 6.1'de normalize edilmiş MAE skorları çizgi şeklinde gösterilmiştir. Farklı diller için normalize edilmiş skor değerleri birbirine yakın olup, MAE skorlarına göre daha düşük varyansa sahiptir. En düşük normalize edilmiş MAE skorları Danca, Flemenkçe ve Almanca'da görülmektedir. Bu üç dilin aynı dil ailesine dahil olması benzer performans göstermelerinde etkili olduğu söylenebilir.

Tek Teklif problemindeki farklı dillerdeki model performansları Şekil 6.2'de gösterilmiştir. En yüksek skorlar Lehçe, Hırvatça ve Romence'de gözlenirken, en düşük skorlar Flemenkçe, Danca, İngilizce ve Fince'de gözlemlenmiştir. Teklif Sayısı problemine benzer şekilde ekonomik olarak gelişmiş ülke dillerindeki performanslar daha düşüktür. Bunun sebebi ekonomik olarak daha az gelişmiş ülkelerde tek teklif alan ihale sayısının fazla olması gösterilebilir. Bu etkiyi incelemek amacıyla Şekil 6.2'de



Şekil 6.1: Teklif Sayısı problemi için test kümesindeki her bir dildeki MBERT performanslarının gösterimi. Çizgi ile her bir dil için ortalama teklif sayısı ile normalize edilmiş MAE skorları gösterilmektedir.

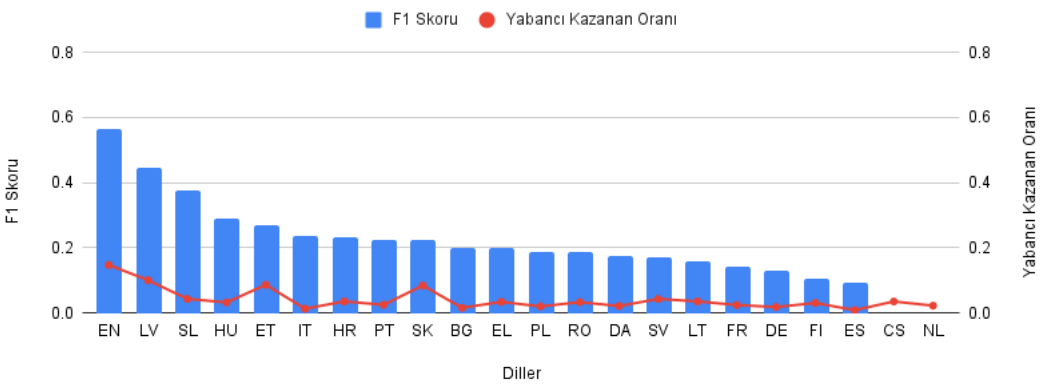
çizgi olarak tek teklif alan ihale oranları gösterilmiştir. Farklı diller için model performanslarının tek teklif oranlarına paralellik gösterdiği söylenebilir.



Şekil 6.2: Tek Teklif problemi için test kümesindeki her bir dildeki MBERT performanslarının gösterimi. Çizgi ile her bir dil için test kümesindeki tek teklifli ihalelerin oranı gösterilmektedir.

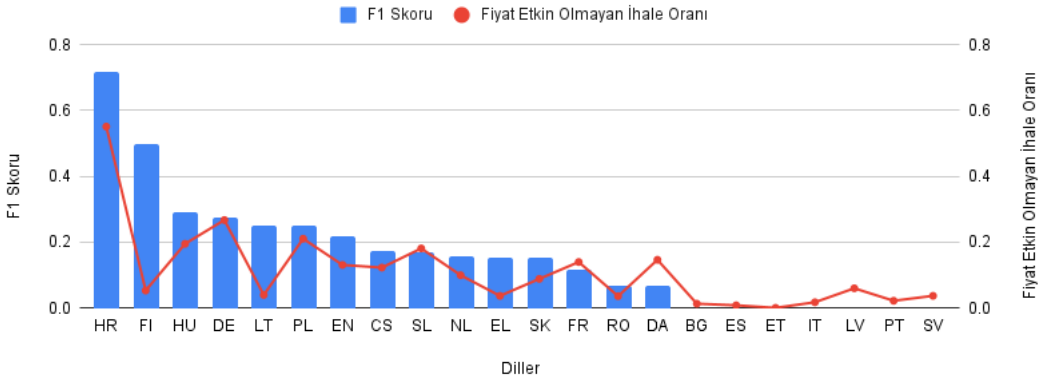
Yabancı Firma problemindeki farklı dillerdeki model performansları Şekil 6.3'te gösterilmiştir. En yüksek skorlar İngilizce, Litvanca, Slovakça, Macarca ve Estonca'da gözlenirken, en düşük skorlar Flemenkçe, Çekçe, Estonca, Fince ve Almanca'da gözlemlenmiştir. En iyi performansı gösteren diller için Tek Teklif problemine benzer şekilde pozitif etiket olan yabancı firma tarafından kazanılan ihale oranlarının etkili olduğu söylenebilir. Bu etki incelemek amacıyla Şekil 6.3'te çizgi olarak yabancı firma tarafından kazanılan ihale oranları gösterilmiştir. Çekçe ve Flemenkçe'de model çoğunluk sınıfı tahmin etmesiyle F_1 0 skorunu vermektedir.

Fiyat Etkin problemindeki farklı dillerdeki model performansları Şekil 6.4'te gösterilmiştir. Yabancı Firma problemine benzer olarak pozitif etiket olan yabancı firma tarafından kazanılan ihale oranlarının etkili olduğu söylenebilir. Fakat Fince ve Litvanca'da pozitif etiket oranının düşük olmasına rağmen model performansının yüksek olduğu söylenebilir. Slovakça, Portekizce, Litvanca, İtalyanca ve İspanyolca'da model çoğunluk sınıfı tahmin etmesiyle F_1 0 skorunu vermektedir. Ayrıca Estonca için



Şekil 6.3: Yabancı Firma problemi için test kümesindeki her bir dildeki MBERT performanslarının gösterimi. Çizgi ile her bir dil için test kümesindeki yabancı firmaların kazandığı ihalelerin oranı gösterilmektedir.

test kümesinde pozitif örnek bulunmamasıyla performansı 0 olarak ölçülmektedir. Sınıflandırma problemleri için genel olarak dil bazındaki performansın o dildeki etiket oranlarına göre değişkenlik göstermesi modellerin farklı dillerdeki öğrenilen bilgileri diğer dillere ne ölçüde aktarabildiği sorusunu ortaya çıkarmaktadır. Bu etkiyi incelemek için çok dilli eğitim karşısında performansın değişimi ve dil geçişli performans ölçümleri incelenmiştir.



Şekil 6.4: Fiyat Etkin problemi için test kümesindeki her bir dildeki MBERT performanslarının gösterimi. Çizgi ile her bir dil için test kümesindeki fiyat etkin olmayan ihalelerin oranı gösterilmektedir.

6.2.4 Çok dilli eğitimin performans üzerine etkisi

Eğitim kümesinde farklı oranlarda farklı dillerden örneklerin bulunması bütün eğitim kümesinin kullanılmasıyla eğitilen modelin performansını etkilemiş olabilir. Bu etkiyi ölçmek amacıyla MBERT dil modeli aşağıdaki 4 şekilde eğitilerek sonuçları karşılaştırılmıştır:

1. **Bütün Diller:** Eğitim kümesinin bütünü ile yani bütün dilleri kapsayacak şekilde hassas ayarlamak. Önceki bölümlerdeki modeller bu şekilde hassas ayarlanmıştır.

2. **Sadece Orjinal Dil:** Her bir dil için eğitim kümesinde o dilde bulunan metinleri kullanarak ayrı bir modeli hassas ayarlamak.

3. **Orjinal Dil + İngilizce:** Her bir dil için eğitim kümesinde o dilde bulunan metinleri ve ek olarak İngilizce örnekleri kullanarak ayrı bir modeli hassas ayarlamak. İngilizce için hassas ayarlama yapılırken örnekler sadece tek bir sefer kullanılmıştır.

4. **Sadece İngilizce:** Eğitim kümesinde sadece İngilizce olan örnekleri kullanarak tek bir modeli hassas ayarlamak.

2 ve 3 numaralı konfigürasyonlar için test kümesi dillere ayrılmış ve ilgili dilde hassas ayarlanan model kullanılarak tahminler oluşturulup bu tahminlerin birleştirilmesiyle bütün test kümesindeki performans hesaplanmıştır.

Çizelge 6.9: Çok dilli eğitimin performans üzerine etkisinin 4 farklı dil kombinasyonu kulanılarak gösterilmesi. En iyi sonuçlar her bir problem için **kalın** ile gösterilmiştir.

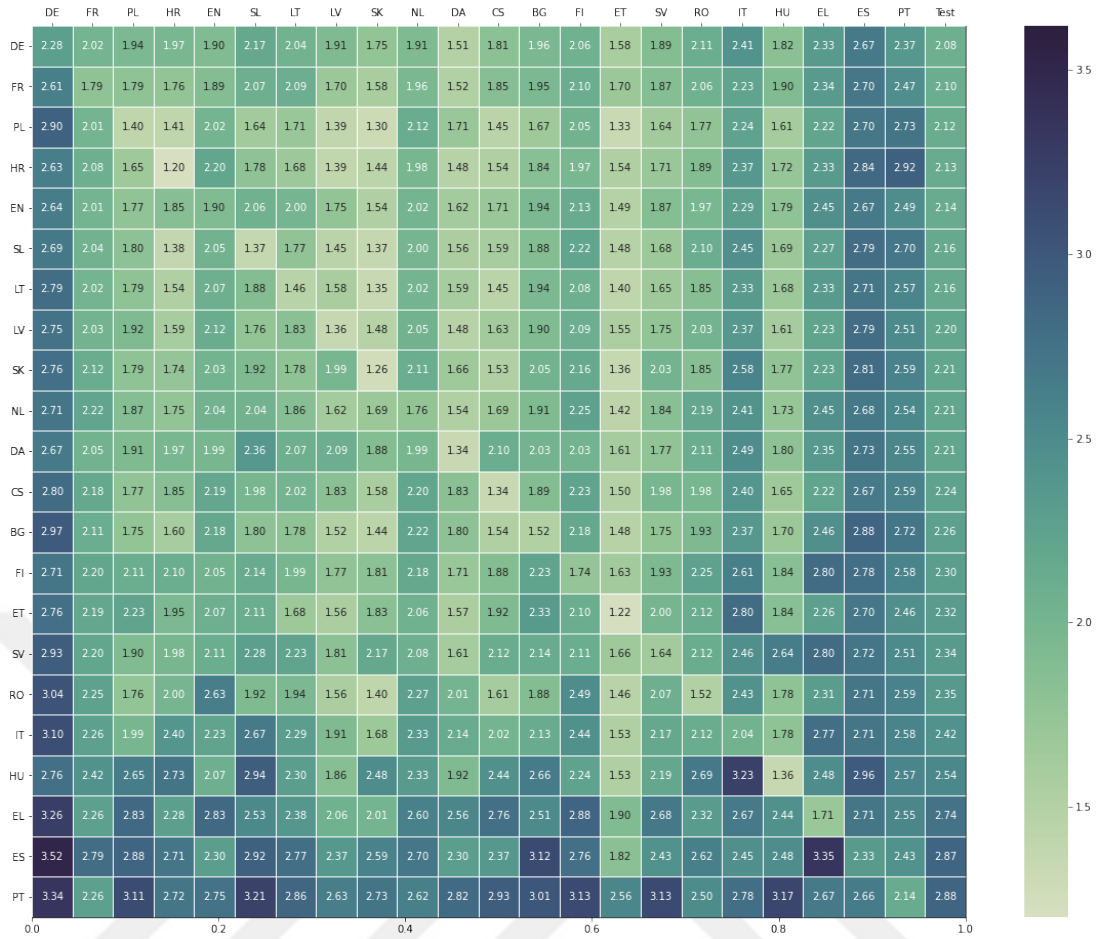
	Tek Teklif	Teklif Sayısı	Yabancı Firma	Fiyat Etkin
Bütün Diller	0.436	1.842	**0.394**	**0.175**
Orjinal Dil + İngilizce	**0.442**	1.840	0.358	0.140
Sadece Orjinal Dil	0.423	**1.830**	0.357	0.143
Sadece İngilizce	0.237	2.139	0.218	0.018

Yapılan deneylerde yalnızca İngilizce olan metinlerle eğitmenin orjinal dilde eğitime göre kazanç sağlamadığı görülmektedir. Sadece Tek Teklif probleminde orijinal dile İngilizce örneklerin eklenmesiyle yapılan hassas ayarlamanın model performansını artırdığı gözlenmektedir. Tek teklif sayısı probleminde de İngilizce ile orjinal dilde model eğitilmesi bütün diller kullanılarak yapılan eğitime göre daha iyi sonuç vermektedir. Pozitif etiket oranlarının görece daha az olduğu sınıflandırma problemlerinde ise bütün dillerde eğitim performansa önemli ölçüde katkı sağlamıştır. Bu da modelin farklı dillerde öğrendiği bilgileri diğer dillerde kullanabilmesini sağladığını göstermektedir.

6.2.5 Dil geçişli eğitimin performans üzerine etkisi

Deneylerin bu kısmında bir önceki kısımdan farklı olarak hangi diller ile eğitimin hangi diller üzerinde etkili olduğu incelenmektedir. Bunun için her bir dil ikilisinin (kaynak-hedef) eğitim-değerlendirme sonuçları karşılaştırılmaktadır. Kaynak dil için o dildeki eğitim kümesindeki örnekler kullanılarak bir model hassas ayarlanır ve test kümesindeki her bir hedef dil için performansı hesaplanır. Genel performansı ise her bir hedef dildeki tahminler birleştirilerek elde edilir.

Teklif Sayısı problemi için dil geçişli deney sonuçları Şekil 6.5'de verilmiştir. Kaynak dil bazında bakıldığında, 22 adet dilden 10 tanesinde en iyi sonuçlar modeller kendi dilindeki örneklerle değerlendirildiğinde alınmaktadır. Hedef dil bazında bakıldığında ise 22 dilden 20 tanesinde en iyi sonuçlar modeller değerlendirilen dilde hassas ayarlandığında alınmıştır. Bu sonuç, aynı dilde örnekleri tahmin etmek için o dildeki

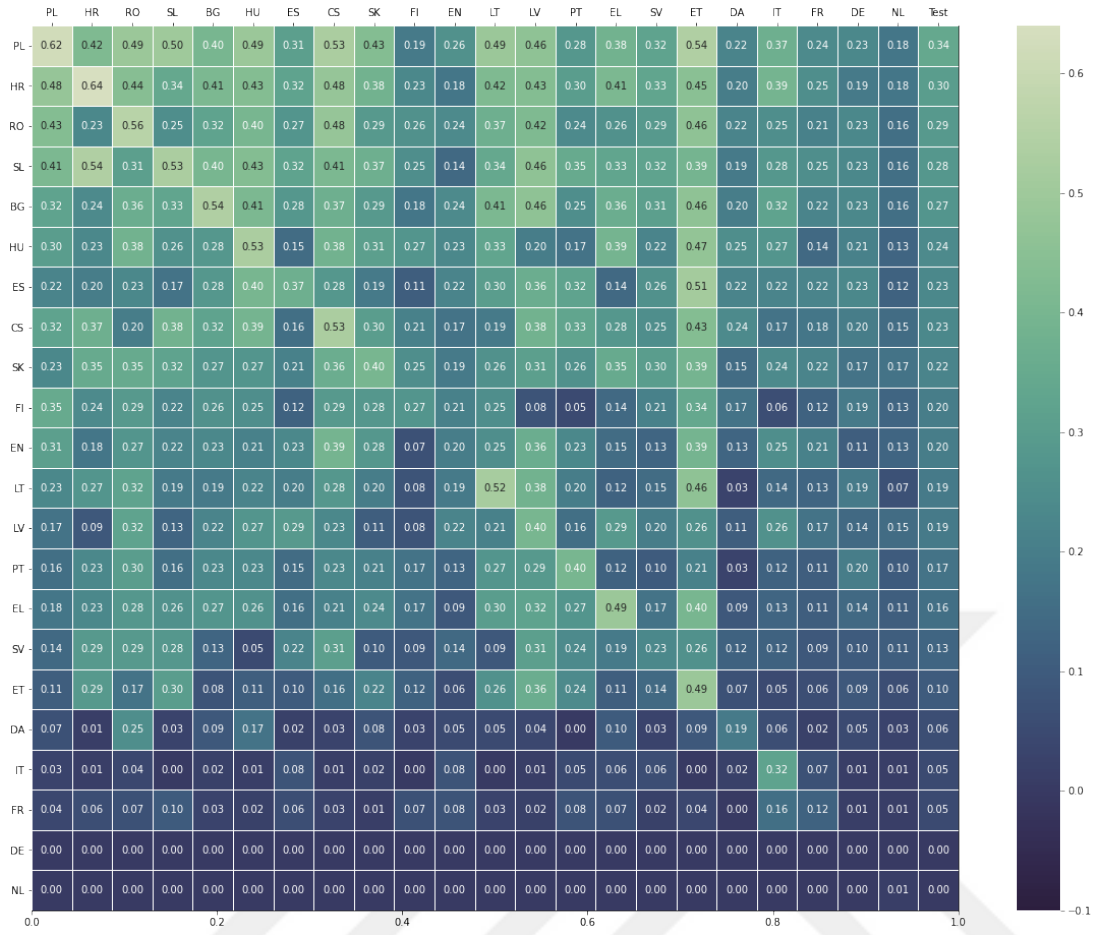


Şekil 6.5: Teklif Sayısı problemi için dil geçişli sonuçların ısı haritası gösterimi. Açık renkler daha düşük MAE skorlarını temsil etmektedir. Satırlar son sütün olan bütün test kümesi performansına göre sıralanmıştır.

örneklerle hassas ayarlamanın iyi bir strateji olduğunu fakat aynı zamanda farklı dillerde de iyi sonuçlar verebileceğini göstermektedir. Yalnızca Portekizce dilindeki örneklerle eğitim yapıldığında bütün test setinde en kötü sonuçlar alınmaktadır. Bunun bir sebebi eğitim kümesinde Portekizce bulunan örneklerin az miktarda olması olabilir.

Test kümesinde en iyi sonuçlar sırasıyla ise yalnız Almanca, Fransızca ve Lehçe olan örneklerle eğitim yapıldığında görülmektedir. Fakat İspanyolca, İtalyanca ve Romence'nin de eğitim kümesinde fazla örneğe sahip olmasına rağmen düşük performans göstermektedir. Bu nedenle örnek sayısının performans üzerinde doğrusal bir etkisi olduğundan bahsetmek mümkün değildir. Ayrıca bütün kaynak hedef ikilileri arasında en iyi sonuçlar sırasıyla Hırvatça, Slovakça, Estonca için alınmıştır. Bu diller de eğitim kümesinde az örneklere sahip dillerdir. Bu ise Şekil 6.1'de de görüldüğü gibi test kümesindeki teklif sayısının dağılımından ileri gelmektedir.

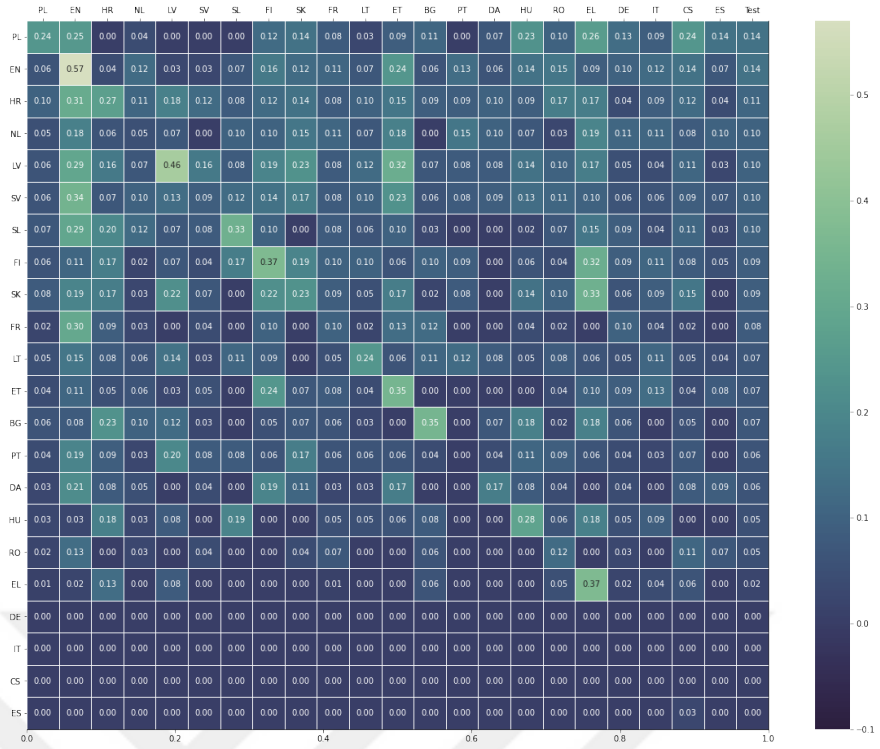
Tek Teklif problemi için ise sonuçlar Şekil 6.6'da gösterilmiştir. Kaynak dil bazında bakıldığında, 22 adet dilden 13 tanesinde en iyi sonuçlar modeller kendi dilindeki örneklerle değerlendirildiğinde alınmaktadır. Hedef dil bazında bakıldığında ise 22 dilden 10 tanesinde en iyi sonuçlar modeller değerlendirilen dilde hassas ayarlandığında alınmıştır. Test kümesindeki Lehçe, Hırvatça ve Macarca eğitilen



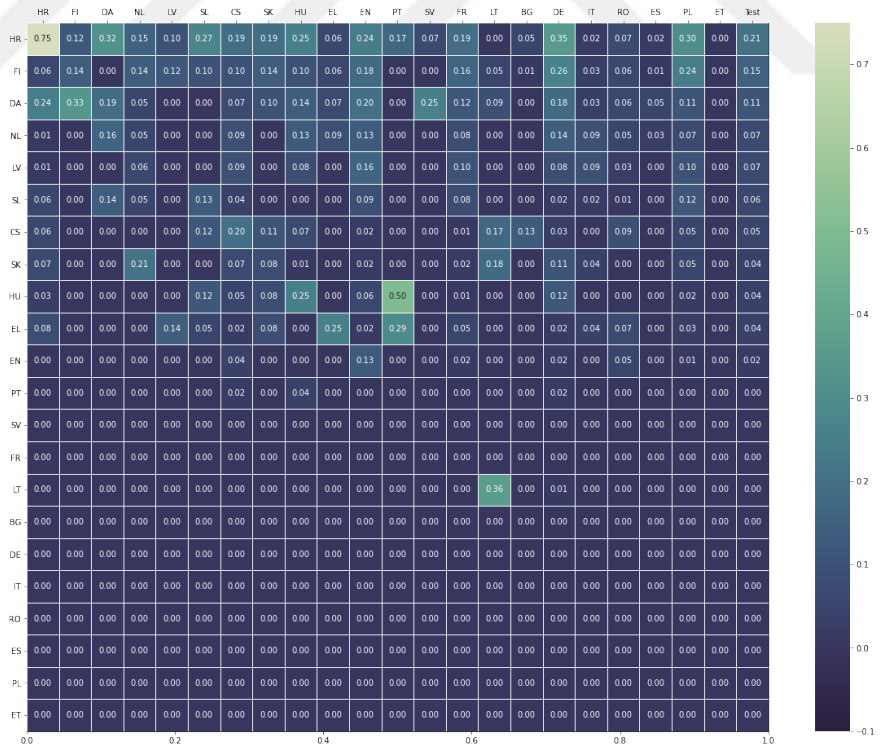
Şekil 6.6: Tek Teklif problemi için dil geçişli sonuçların ısı haritası gösterimi. Açık renkler daha yüksek F_1 skorlarını temsil etmektedir. Satırlar son sütün olan bütün test kümesi performansına göre sıralanmıştır.

modeller 13 farklı dilde en iyi sonuçları vermiştir. Buradan hareketle Teklif Sayısı problemine göre bu problem için dil geçişliliğinin daha yüksek olduğu söylenebilir. Almanca ve Flemenkçe dillerinde eğitilen modeller bütün dillerde 0 F_1 skoru vermektedir. En iyi modeller ise Lehçe, Hırvatça, Romence, Slovakça olarak pozitif etiket oranıyla paralellik göstermektedir.

Yabancı Firma ve Fiyat Etkin problemleri için ise sonuçlar sırasıyla Şekil 6.8'de ve Şekil 6.7'de gösterilmiştir. 22 adet dilden 9 tanesinde en iyi sonuçlar modeller kendi dilindeki örneklerle değerlendirildiğinde alınırken 12 tanesinde en iyi sonuçlar modeller değerlendirilen dilde hassas ayarlandığında alınmıştır. Kaynak hedef ikilileri açısından karşılaştırıldığında en yüksek performansı İngilizce - İngilizce ikilisi göstermiştir. Ayrıca 6 farklı dilde en iyi model performansı İngilizce örnekler üzerinde elde edilmiştir. Bunun sebebi Şekil 6.3'te de görülebileceği gibi pozitif örnek sayısından kaynaklıdığı söylenebilir. Almanca, İtalyanca, Çekçe ve İspanyolca'da ise modeller meydan değeri tahmin ederek bütün dillerde 0 F_1 skoru vermektedirler. Fiyat Etkin probleminde en iyi sonuçlar pozitif etiket oranının en çok olduğu Hırvatça'da alınmıştır. Ayrıca birçok dilde modellerin ilgili hedef ölçütü öğrenemediği görülmektedir. Fakat Şekil 6.4'teki sonuçlarla karşılaştırıldığında çok dilli eğitimin birçok dilde öğrenmeyi gerçekleştirdiği görülebilir.



Şekil 6.7: Yabancı Firma problemi için dil geçişli sonuçların ısı haritası gösterimi. Açık renkler daha yüksek F_1 skorlarını temsil etmektedir. Satırlar son sütün olan bütün test kümesi performansına göre sıralanmıştır.



Şekil 6.8: Fiyat Etkin problemi için dil geçişli sonuçların ısı haritası gösterimi. Açık renkler daha yüksek F_1 skorlarını temsil etmektedir. Satırlar son sütün olan bütün test kümesi performansına göre sıralanmıştır.

7. SONUÇ VE ÖNERİLER

Bu çalışmada kamu alımı ilan açıklamalarını kullanarak Avrupa Birliği kamu ihalelerindeki rekabetin ve maliyet etkinliğinin nasıl tahmin edilebileceği araştırılmıştır. Bunun için 1) teklif sayısı, 2) ihalenin tek teklif alıp almadığı 3) ihale edilen firmanın yabancı olup olmadığı ve 4) sözleşme fiyatının tahmin edilen maliyet fiyatını aşıp aşmadığı olmak üzere dört tahmin senaryosu incelenmiştir. AB tarafından paylaşılan TED verisinden oluşturduğumuz veri kümesi 22 farklı dili kapsamasıyla ve oldukça dengesiz etiket dağılımına sahip olması ile araştırmada çeşitli zorlukları beraberinde getirmiştir. Çok dilli dönüştürücü modellerinin hassas ayarlanması, sektör tabanlı filtrelemeyle model performanslarının değişimi gibi farklı yöntemler karşılaştırılmıştır. Kapsamlı deneyler sonucunda aşağıdaki gözlemlere ulaşılmıştır.

İlk olarak, önerilen modeller tüm tahmin görevlerinde temel modellerden daha iyi performans göstermiştir. Bu durum kamu ilan metinlerinin ihale sonuçları üzerinde etkili olduğunu göstermektedir. Fakat unutulmamalıdır ki, raporlanan en iyi sonuçlar kurulu deney düzeneklerinde ve sadece metinleri kullanarak elde edilmektedir. İhale süreçlerindeki diğer etkenler kullanılarak model performanslarının geliştirilmesi mümkündür.

İkinci olarak, teklif sayısı, tek teklifli ihaleler ve yabancı firma kazanımı tahmini görevlerinde MBERT hassas ayarlanan model en iyi performansı göstermiştir. MBERT cümle gösterim vektörlerinin kullanıldığı MLP modeli ise maliyet etkinliği tahmin probleminde en iyi sonucu vermiştir. Hassas ayarlanan modellerden MBERT her dört görevde de XLMR'dan daha iyi sonuç vermektedir.

Üçüncü olarak, çok dilli eğitim orjinal dildeki eğitimden daha yüksek tahmin performansı göstermiştir. Özellikle etiket oranının görece az olduğu Yabancı Firma ve Fiyat Etkinliği problemlerinde diğer dillerde öğrenilen bilginin başka dillerde yapılan tahminlerde kullanılabildiği görülmektedir.

Son olarak, diller arası eğitimde dil bazındaki etiket dağılımı modellerimizin performansı üzerinde etkili olmaktadır. Sınıflandırma problemlerinde dil geçişliliği regresyon problemine göre daha yüksek gözlemlenmektedir.

Bu çalışmanın devamı olarak gelecekteki araştırmalar şu yönlerde devam edebilir: İlk olarak, diller arasında dengesiz etiket dağılımının olumsuz etkisinin nasıl azaltılacağı üzerinde çalışılabilir. İkinci olarak, yalnız açıklama kısımlarını kullanmak yerine ihale ilan dokümanlarının bütünü kullanılabilir. Ek 1'de görülebileceği gibi ihale metinleri açıklamaların yanı sıra ihale kriterleri ve kontrat süreleri gibi ihale sonuçları üzerinde etkili olabilecek önemli bilgileri de içermektedir. Üçüncü olarak, çok dilli eğitimin tahmin doğruluğunu geliştirdiğini gösterdiğimiz için, AB dışı ülkelerin kamu ihale çağrıları veri kümesinin genişletilerek modellerin iyileştirilmesi için kullanılabilir.

Son olarak, rekabet gücünü ve maliyet etkinliğini artırmak için bir kamu alım ilan metni açıklamasının nasıl değiştirileceğini önermek amacıyla GPT gibi metin oluşturma modelleri kullanılarak önerme modelleri geliştirilebilir.

KAYNAKLAR

[1] **Ahmia, O., Béchet, N., and Marteau, P.-F.** Two multilingual corpora extracted from the tenders electronic daily for machine learning and machine translation applications. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018).

[2] **Bosio, E., and Djankov, S.** How large is public procurement? *World Bank Blogs* (2020).

[3] **Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V.** Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020* (2020), D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds., Association for Computational Linguistics, pp. 8440–8451.

[4] **Coviello, D., Guglielmo, A., and Spagnolo, G.** The effect of discretion on procurement performance. *Management Science 64*, 2 (2018), 715–738.

[5] **Deltas, G., and Evenett, S.** Language as a barrier to entry: Foreign competition in georgian public procurement. *International Journal of Industrial Organization* (2020).

[6] **Demiray, Y., Ozbayoglu, A., and Tas, B.** Estimation of the number of participants in government tenders with computational intelligence. In *3rd Workshop on Social and Algorithmic Issues in Business Support (SAIBS)* (2015), pp. 433–437.

[7] **Devlin, J., Chang, M., 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (2019), J. Burstein, C. Doran, and T. Solorio, Eds., Association for Computational Linguistics, pp. 4171–4186.

[8] **Estache, A., and Iimi, A.** Bidder asymmetry in infrastructure procurement: Are there any fringe bidders? *Review of Industrial Organization 36* (2010), 163–187.

[9] **Fazekas, M., and Tóth, B.** The extent and cost of corruption in transport infrastructure. new evidence from europe. *Transportation research part A: policy and practice 113* (2018), 35–54.

[10] **from the Commission, C.** Directive 2014/24/eu of the european parliament and of the council of 26 february 2014 on public procurement and repealing directive 2004/18/ec". *Official Journal of the European Union* (2014).

[11] **García Rodríguez, M. J., Rodríguez Montequín, V., Ortega Fernández, F., and Villanueva Balsera, J. M.** Public procurement announcements in spain: regulations, data analysis, and award price estimator using machine learning. *Complexity 2019* (2019).

[12] **Gorgun, M. K., Kutlu, M., and Taş, B. K. O.** Predicting the number of bidders in public procurement. In *2020 5th International Conference on Computer Science and Engineering (UBMK)* (2020), IEEE, pp. 360–365.

[13] **Iimi, A.** Auction reforms for effective official development assistance. *Review of Industrial Organization 28* (2006), 109–128.

[14] **Ishii, R.** Favor exchange in collusion: Empirical study of repeated procurement auctions in japan. *International Journal of Industrial Organization 27*, 2 (2009), 137–144.

[15] **Jang, Y., Yi, J., Son, J., Kang, H., and Lee, J.** Development of classification model for the level of bid price volatility of public construction project focused on analysis of pre-bid clarification document. In *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction* (2019), vol. 36, IAARC Publications, pp. 1245–1253.

[16] **Karthikeyan, K., Wang, Z., Mayhew, S., and Roth, D.** Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations* (2019).

[17] **Kutlina-Dimitrova, Z., and Lakatos, C.** Determinants of direct cross-border public procurement in eu member states. *Review of World Economics 152*, 3 (2016), 501–528.

[18] **Lee, J., and Yi, J.-S.** Predicting project's uncertainty risk in the bidding process by integrating unstructured text data and structured numerical data using text mining. *Applied Sciences 7*, 11 (2017), 1141.

[19] **Lima, M., Silva, R., de Souza Mendes, F. L., de Carvalho, L. R., Araujo, A., and de Barros Vidal, F.** Inferring about fraudulent collusion risk on brazilian public works contracts in official texts using a bi-lstm approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (2020), pp. 1580–1588.

[20] **Mehrbod, A., and Grilo, A.** Tender calls search using a procurement product named entity recogniser. *Advanced Engineering Informatics 36* (2018), 216–228.

[21] **Onur, İ., Özcan, R., and Taş, B. K. O.** Public procurement auctions and competition in turkey. *Review of industrial organization 40*, 3 (2012), 207–223.

[22] **Onur, I., Ozcan, R., and Tas, B. K. O.** Public procurement auctions and competition in turkey. *Review of Industrial Organization 40* (2012), 207–223.

[23] **Onur, I., and Tas, B. K. O.** Optimal bidder participation in public procurement auctions. *International Tax and Public Finance 26* (2019), 595–617.

[24] **Pappas, N., and Popescu-Belis, A.** Multilingual hierarchical attention networks for document classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (2017), pp. 1015–1025.

[25] **Peña-López, I., et al.** Government at a glance 2019.

[26] **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.

[27] **Rabuzin, K., and Modrusan, N.** Prediction of public procurement corruption indices using machine learning methods. In *KMIS* (2019), pp. 333–340.

[28] **Simperl, E., Corcho, O., Grobelnik, M., Roman, D., Soylu, A., Ruíz, M. J. F., Gatti, S., Taggart, C., Klima, U. S., Uliana, A. F., et al.** Towards a knowledge graph based platform for public procurement. In *Research Conference on Metadata and Semantics Research* (2018), Springer, pp. 317–323.

[29] **Tas, B. K. O., Dawar, K., Holmes, P., and Togan, S.** Does the wto government procurement agreement deliver what it promises? *World Trade Review 18*, 4 (2019), 609–634.

[30] **Tas, B. K. O., D. K. H. P. . T. S.** Does the wto government procurement agreement deliver what it promises? *World Trade Review 18*, 4 (2019), 609–634.

[31] **the World Bank MENA Procurement Team**. Why reform public procurement? *The Middle East and North Africa (MENA) regional procurement conference* (2012).

[32] **van der Heijden, N., Yannakoudakis, H., Mishra, P., and Shutova, E.** Multilingual and cross-lingual document classification: A meta-learning approach. *arXiv preprint arXiv:2101.11302* (2021).

[33] **Williams, T. P., and Gong, J.** Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Automation in Construction 43* (2014), 23–29.

[34] **Wu, S., and Dredze, M.** Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference*

on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 833–844.

EKLER

EK 1 : PDF formatında ihale örneği
EK 2 : XML formatında ihale örneği
EK 3 : TED ihale dokümanı çeşitleri
EK 4 : Veri kümesindeki diller ülkeler ve karşılık gelen kodlar
EK 5 : Spacy POS etiket açıklamaları
EK 6 : Spacy NER etiket açıklamaları
EK 7 : Diller için Spacy NER etiketleri farklılıkları

EK 1

This notice in TED website: https://ted.europa.eu/udl?uri=TED:NOTICE:191599-2018:TEXT:EN:HTML

Netherlands-Delfzijl: Education and training services
2018/S 085-191599

Contract notice

Services

Legal Basis:
Directive 2014/24/EU

Section I: Contracting authority

I.1) **Name and addresses**
Official name: Eemsdeltacollege
National registration number: 36788065
Postal address: Sikkel 3
Town: Delfzijl
NUTS code: NL NEDERLAND
Postal code: 9932 BD
Country: Netherlands
Contact person: Marleen Kempa
E-mail: info@eemsdeltacollege.nl
Telephone: +31 596693693
Fax: +31 596693678
Internet address(es):
Main address: http://www.eemsdeltacollege.nl

I.3) **Communication**
The procurement documents are available for unrestricted and full direct access, free of charge, at: https://www.tenderned.nl/tenderned-web/aankondiging/detail/publicatie/akid/8b8909460ed702da29db13afb3ab30fb
Additional information can be obtained from the abovementioned address
Tenders or requests to participate must be submitted electronically via: https://www.tenderned.nl/tenderned-web/aankondiging/detail/publicatie/akid/8b8909460ed702da29db13afb3ab30fb
Tenders or requests to participate must be submitted to the abovementioned address

I.4) **Type of the contracting authority**
Body governed by public law

I.5) **Main activity**
Education

Section II: Object

II.1) **Scope of the procurement**

II.1.1) **Title:**
Best Value aanbesteding ICT-Onderwijs
Reference number: 2018-03

II.1.2) **Main CPV code**
80000000 Education and training services - FA01

II.1.3) **Type of contract**

Şekil 7.1: Örnek bir ihale dokümanı

Services

II.1.4) **Short description:**
Het leveren en implementeren van een Leerlijn ter stimulering van de digitale geletterdheid van leerlingen, aangeboden aan leerkrachten/docenten. Deze leerlijn is vraag gestuurd, gericht op het Primair Onderwijs en Voortgezet Onderwijs en wordt met en door de leerkracht/docent uitgevoerd op de school. De leerlijn is in de basis geïntegreerd met het bestaande onderwijs.

II.1.5) **Estimated total value**
Value excluding VAT: 3 572 000.00 EUR

II.1.6) **Information about lots**
This contract is divided into lots: no

II.2) **Description**

II.2.3) **Place of performance**
NUTS code: NL NEDERLAND
NUTS code: NL1 NOORD-NEDERLAND

II.2.4) **Description of the procurement:**
De omvang van de opdracht ligt op 3 572 000 EUR inclusief BTW. De omvang betreft de scholen in het bevingengebied.

II.2.5) **Award criteria**
Criteria below
Quality criterion - Name: Prestatie Onderbouwing met planning / Weighting: 30
Quality criterion - Name: Risicodossier / Weighting: 20
Quality criterion - Name: Kansendossier / Weighting: 15
Quality criterion - Name: Interviews / Weighting: 35
Cost criterion - Name: Prijs / Weighting: 0

II.2.6) **Estimated value**
Value excluding VAT: 3 572 000.00 EUR

II.2.7) **Duration of the contract, framework agreement or dynamic purchasing system**
Start: 03/09/2018
End: 29/08/2022
This contract is subject to renewal: yes
Description of renewals:
De opdracht kan worden verlengd met tweemaal (2) een (1) jaar.

II.2.10) **Information about variants**
Variants will be accepted: no

II.2.11) **Information about options**
Options: yes
Description of options:
De opties betreffen verlengingen.

II.2.13) **Information about European Union funds**
The procurement is related to a project and/or programme financed by European Union funds: no

II.2.14) **Additional information**

Section III: Legal, economic, financial and technical information

III.1) **Conditions for participation**

III.1.1) **Suitability to pursue the professional activity, including requirements relating to enrolment on professional or trade registers**

List and brief description of conditions:
— Geen crimineel verleden,
— Geen ernstige beroepsfout,
— Geen vervalsing mededinging,
— Betalingen belastingen en premies,
— Geen faillissement of surseance van betaling,
— Geen gerechtelijke uitspraak,
— Geen vervalsing mededinging,
— Geen betrokkenheid voorbereiding,
— Geen valse verklaringen.

III.1.2) **Economic and financial standing**

List and brief description of selection criteria:
— Financiële gegoedheid,
— Bedrijfs- en beroepsaansprakelijkheid.

III.1.3) **Technical and professional ability**

List and brief description of selection criteria:
— Technische en/of beroepsbekwaamheid.

Section IV: Procedure

IV.1) **Description**

IV.1.1) **Type of procedure**

Open procedure

IV.1.3) **Information about a framework agreement or a dynamic purchasing system**

IV.1.8) **Information about the Government Procurement Agreement (GPA)**

The procurement is covered by the Government Procurement Agreement: yes

IV.2) **Administrative information**

IV.2.2) **Time limit for receipt of tenders or requests to participate**

Date: 19/06/2018
Local time: 12:00

IV.2.3) **Estimated date of dispatch of invitations to tender or to participate to selected candidates**

IV.2.4) **Languages in which tenders or requests to participate may be submitted:**

Dutch

IV.2.6) **Minimum time frame during which the tenderer must maintain the tender**

Tender must be valid until: 01/09/2018

IV.2.7) **Conditions for opening of tenders**

Date: 19/06/2018
Local time: 12:05

Section VI: Complementary information

VI.1) **Information about recurrence**

This is a recurrent procurement: no

VI.3) **Additional information:**

OP 14.5.2018 is er de gelegenheid om de marktbijeenkomst ten aanzien van de Best Value Approach bij te wonen. Indien u interesse heeft in deze bijeenkomst kunt u zich uiterlijk voor 11.5.2018 (12:00) aanmelden door een bericht te sturen via Tenderned.
NB: voor deze opdracht treedt Eemsdeltacollege op als penvoerder voor de vereniging "Kansrijke Groningers" in oprichting. Deze vereniging bedient meerdere scholen in het bevingengebied, voor nadere toelichting zie het Beschrijvend document.

VI.4) **Procedures for review**

VI.4.1) **Review body**

Official name: Rechtbank Noord-Nederland/ locatie Groningen
Town: Groningen
Country: Netherlands
E-mail: info@rechtspraak.nl
Internet address: http://www.rechtspraak.nl

VI.4.3) **Review procedure**

Precise information on deadline(s) for review procedures:
Zie Beschrijvend document.

VI.5) **Date of dispatch of this notice:**

01/05/2018

EK 2

Şekil 7.2: Örnek bir kamu alım ilan dokümanının XML formatında görünümü. XML ağacının son yaprağı olarak sadece çalışmada kullanılan metinleri içeren SHORT_DESCR kısmı gösterilmiştir.

EK 3

Çizelge 7.1: TED ihale doküman çeşitleri listesi

0	Prior information notice without call for competition
1	Corrigendum
2	Additional Information
3	Contract Notice
4	Prequalification Notices
7	Contract Award Notice
A	Prior information notice with call for competition
B	Buyer profile
C	Works concession
D	Design contest
E	Works contracts awarded by the concessionnaire
F	Subcontracts in the fields of defence and security
G	European economic interest grouping
H	Services concession
M	Periodic indicative notice with call for competition
O	Qualification system with call for competition
P	Periodic indicative notice without call for competition
Q	Qualification system without call for competition
R	Results of design contests
S	European company
V	Voluntary ex ante transparency notice
Y	Dynamic purchasing system

EK 4

Çizelge 7.2: Veri kümesindeki diller ve karşılık gelen kodları

Dil Kodu	**Dil İsmi**	**Spacy**	**Açıklama Etiketi**
DA	Danish	✓	Beskrivelse
NL	Dutch	✓	Beschrijving
EN	English	✓	Description
FR	French	✓	Description
DE	German	✓	Beschreibung
EL	Greek	✓	´
IT	Italian	✓	Descrizione
LT	Lithuanian	✓	apibūdinimas
PL	Polish	✓	Opis
PT	Portuguese	✓	Descrição
RO	Romanian	✓	Descrierea
ES	Spanish	✓	Descripción
BG	Bulgarian	✗	
FI	Finnish	✗	kuvaus
SV	Slovak	✗	Beskrivning
LV	Latvian	✗	apraksts
CS	Czech	✗	Popis
SK	Slovak	✗	Opis
HR	Crotian	✗	Opis
SL	Slovenian	✗	Opis
ET	Estonian	✗	kirjeldus
HU	Hungarian	✗	meghatározása
GA	Irish	✗	Cur síos ar
MT	Maltese	✗	deskrizzjoni

Çizelge 7.3: Veri kümesinde bulunan ülkeler ve ISO kodları

ISO Kodu	**Ülke**
DE	Germany
FR	France
PL	Poland
UK	United Kingdom
NL	Netherlands
IT	Italy
ES	Spain
RO	Romania
LT	Lithuania
CZ	Czech Republic
NO	Norway
BG	Bulgaria
SE	Sweden
HU	Hungary
AT	Austria
DK	Denmark
CH	Switzerland
BE	Belgium
HR	Croatia
FI	Finland
SK	Slovakia
IE	Ireland
GR	Greece
LV	Latvia
EE	Estonia
SI	Slovenia
PT	Portugal
LU	Luxembourg
MT	Malta
CY	Cyprus
MK	Makedonia
IS	Iceland
LI	Liechtenstein

EK 5

Çizelge 7.4: Spacy POS etiketleri

ADJ	adjective
ADP	adposition
ADV	adverb
AUX	auxiliary
CONJ	conjunction
CCONJ	Coordinating, conjunction
DET	determiner
INTJ	interjection
NOUN	noun
NUM	numeral
PART	particle
PROPN	proper noun
PRON	pronoun
PUNCT	punctuation
SCONJ	subordinating conjunction
SYM	symbol
VERB	verb
X	other
SPACE	space

EK 6

Çizelge 7.5: Spacy NER etiket açıklamaları

LOC	Non-GPE locations, mountain ranges, bodies of water
MISC	Miscellaneous entities, e.g. events, nationalities, products or works of art
ORG	Companies, agencies, institutions, etc.
PER	Named person or family.
CARDINAL	Numerals that do not fall under another type
DATE	Absolute or relative dates or periods
EVENT	Named hurricanes, battles, wars, sports events, etc.
FAC	Buildings, airports, highways, bridges, etc.
GPE	Countries, cities, states
LANGUAGE	Any named language
LAW	Named documents made into laws.
MONEY	Monetary values, including unit
NORP	Nationalities or religious or political groups
ORDINAL	"first", "second", etc.
PERCENT	Percentage, including "%"
PERSON	People, including fictional
PRODUCT	Objects, vehicles, foods, etc. (not services)
QUANTITY	Measurements, as of weight or distance
TIME	Times smaller than a day
WORK_OF_ART	Titles of books, songs, etc.

EK 7

Çizelge 7.6: Her bir dil için Spacy NER etiketleri

DA	LOC, MISC, ORG, PER
NL	CARDINAL, DATE, EVENT, FAC, GPE, LANGUAGE, LAW, LOC, MONEY, NORP, ORDINAL, ORG, PERCENT, PERSON, PRODUCT, QUANTITY, TIME, WORK_OF_ART
EN	CARDINAL, DATE, EVENT, FAC, GPE, LANGUAGE, LAW, LOC, MONEY, NORP, ORDINAL, ORG, PERCENT, PERSON, PRODUCT, QUANTITY, TIME, WORK_OF_ART
FR	LOC, MISC, ORG, PER
DE	LOC, MISC, ORG, PER
EL	EVENT, GPE, LOC, ORG, PERSON, PRODUCT
IT	LOC, MISC, ORG, PER
LT	GPE, LOC, ORG, PERSON, PRODUCT, TIME
PL	date, geogName, orgName, persName, placeName, time
PT	LOC, MISC, ORG, PER
RO	DATETIME, EVENT, FACILITY, GPE, LANGUAGE, LOC, MONEY, NAT_REL_POL, NUMERIC_VALUE, ORDINAL, ORGANIZATION, PERIOD, PERSON, PRODUCT, QUANTITY, WORK_OF_ART
ES	LOC, MISC, ORG, PER