

Ön Eğitilmiş Dil Modelleri Kullanarak Türkçe Tweetlerden Cinsiyet Tespiti

İlhami SEL^{1*}, Davut HANBAY²

¹ Bilgisayar Mühendisliği, Mühendislik Fakültesi, İnönü Üniversitesi, Malatya, Türkiye

² Bilgisayar Mühendisliği, Mühendislik Fakültesi, İnönü Üniversitesi, Malatya, Türkiye

*¹ ilhamisel23@gmail.com, ² davut.hanbay@inonu.edu.tr

(Geliş/Received: 28/04/2021;

Kabul/Accepted: 25/08/2021)

Öz: Yazar profili oluşturma (Author Profiling) bir metnin üslup ve içeriğine bakarak yazarın çeşitli özelliklerinin ortaya çıkarılmasına yönelik bir metin kümesi analizidir. Bu özellikler yaş, cinsiyet, kişilik özellikleri ve hatta meslek gibi unsurları barındırır. Cinsiyet belirleme yazar profili oluşturma çalışmalarının alt alanlarından birisidir. Siber suçlar başta olmak üzere sahte haber yayma gibi adli olayların yanında pazarlama (reklamcılık), sosyolojik ve psikolojik olayların incelenmesinde cinsiyet belirleme oldukça önemlidir. Ayrıca İletişim partnerinin cinsiyetini belirlemek, e-posta, bloglar, forumlar gibi sosyal medya aracılığıyla gerçekleşen sahtekarlık ve suistimallerin önlenmesine yardımcı olabilmektedir. Türkçe dili için kısa gönderilerden cinsiyet tespiti yapılması, diğer dillere oranla çok fazla çalışılmayan bir alandır. Bu çalışmada Türkçe Twitter gönderilerinden cinsiyet tespiti yapılmaya çalışılmıştır. Twitter gönderileri dil kurallarına uymayan, kısaltılmış kelimeler ve anlamsız cümle yapıları da içerme ihtimallerine rağmen cinsiyet belirleme görevi için yaygın bir şekilde kullanılmaktadır. Problem bir sınıflandırma görevi olarak ele alınmıştır. Yapılan çalışmada makine öğrenmesi metodları (TF-IDF + SVM), derin öğrenme yöntemleri (LSTM, CNN) ve Türkçe için ön eğitilmiş dil modelleri (BERT, DistilBert, Electra) kullanılmıştır. Yapılan deneyler sonucunda en yüksek başarıyı (%80.1) kelime boyutunun 128k olduğu Bert modeli sağlamıştır. Bu çalışma diğer metin sınıflandırma görevleri için de detaylı bir çalışma olma özelliği göstermektedir.

Anahtar kelimeler: Yazar Profili Oluşturma, Cinsiyet Tespiti, Doğal Dil İşleme, Dil Modelleri, Metin Sınıflandırma.

Gender Identification from Turkish Tweets Using Pre-Trained Language Models

Abstract: Author Profiling is a text set analysis to reveal various characteristics of the author by examining the style and content of a text. These features include factors such as age, gender, personality traits and even profession. Gender identification is one of the subfields of author profile creation. Gender identification is very important in the investigation of marketing (advertising), sociological and psychological events, as well as forensic events such as spreading fake news, especially cybercrime. In addition, identifying the gender of the communication partner can help prevent fraud and abuse through social media such as e-mail, blogs, forums. Identifying the gender of short posts for the Turkish language is an area that has not been studied much compared to other languages. In this study, it was attempted to determine gender from Turkish Twitter posts. Twitter posts are widely used for gender identification, although they may include ungrammatical structures, abbreviated words and meaningless sentence structures. The problem is handled as a classification task. In the study, machine learning methods (TF-IDF + SVM), deep learning methods (LSTM, CNN) and pre-trained language models for Turkish (BERT, DistilBert, Electra) were used. As a result of the experiments, Bert model with the word size of 128k provided the highest success (80.1%). This study also features as a detailed study for other text classification tasks.

Key words: Author Profiling, Gender Identification, Natural Language Processing, Language Models, Text Classification.

1. Giriş

Yazar profili oluşturma görevi için metin kümesinin çıkarılmasında sosyal medya oldukça zengin bir kaynaktır. Başta Twitter olmak üzere çok sayıda sosyal medya uygulaması kişilerin duygu ve düşüncelerini kolayca paylaşabilmelerini sağlamaktadır. Bu günden güne giderek artan paylaşım sayıları sayesinde kişilerin yazar profilleri hakkında belirli bilgilere ulaşmak mümkün olmuştur. Ne yazık ki günlük yaşamın ayrılmaz bir parçası haline gelen sosyal medya araçları her zaman güvenilir ve doğru olmamaktadır. Bu noktada dolandırıcılık, istismar gibi siber suçların tespitinde ve önlenmesinde yazar profili çıkarmanın önemi de artmaktadır. Öte yandan reklam pazarı açısından kullanıcıların tercihleri ve beklentileri pazar arayışı içinde olan firmalar için oldukça önemlidir.

* Sorumlu yazar: ilhamisel@msn.com. Yazarların ORCID Numarası: ¹ 0000-0003-0222-7017, ² 0000-0003-2271-7865

Her yıl düzenlenen uluslararası CLEF¹ (Conference and Labs of the Evaluation Forum) konferanslarında çeşitli PAN² görevleri bulunmaktadır. Yazar profili oluşturma da bu görevler arasındadır. Yaş, cinsiyet, eğitim gibi temel özelliklerin tespiti görevlerinin yanında, herhangi bir yazarın sahte haber yaymaya istekli olup olmadığını belirlenmesi gibi özel görevler de bulunmaktadır [1]. PAN2013 [2, 3] ten itibaren bu görevler içerisine giren cinsiyet tespiti bugün bile önemini korumaktadır. Özellikle siber suçların tespitinde önemli bir yeri olan cinsiyet belirleme görevi İngilizce, İspanyolca gibi dillerde önemli başarılarla ulaşmasına rağmen Türkçe için yeterli sayıda çalışma yapılmamıştır.

Cinsiyet belirleme görevi ikili metin sınıflandırma problemi olarak ele alınmaktadır. Metin sınıflandırma probleminden farklı olarak sosyal medya paylaşımlarıyla sınıflandırma yapmak oldukça karmaşık ve başarısı düşük olabilmektedir. Bu durumun genel sebepleri şu şekilde sıralanabilir:

- Metinlerin bir veya iki kelime gibi küçük boyutlarda olabilmesi
- Kelimelerin sözlükte kullanıldığı şekilde yazılmayıp bazı harflerin eksik veya bazılarının fazla yazılabilmesi
- Metin içinde değinme (Mention), etiket (Hashtag), bağlantı (url) veya sındamga (emoji) gibi web ifadelerin oldukça fazla olabilmesi

Türkçe için Sezerer ve Ark. tarafından [4] hazırlanan ve Twitter'dan toplanmış cinsiyetlerin etiketlendiği veri seti bu çalışma için ulaşabildiğimiz en kapsamlı veri setidir. Ayrıca yazarların bu veri setiyle yaptıkları çalışmada (Bag-Of-Words + SVM) %72.32'lik başarıya ulaşmışlardır. Aynı veri seti bu çalışma için de kullanılmıştır.

Bu çalışma da Türkçe veri setleriyle önceden eğitilmiş BERT [5], Distilbert [6] ve Electra[7] gibi mimarilerde dil modelleri ve türevleri kullanılarak cinsiyet tespiti yapılmaya çalışılmıştır. Bu dil modelleri üzerinde ince ayar (fine tuning) yapılarak metin sınıflandırma görevi için başarımları hesaplanmıştır. Yapılan deneyler sonucunda en yüksek başarı puanını (%80.1) Bert modeli almıştır. Bu çalışma Türkçe dili için yazar profili oluşturma görevlerinden cinsiyet tespiti için sınırlı ve az sayıda yapılmış çalışmalar arasındadır.

Bu çalışmanın ikinci bölümünde ilgili çalışmalar, üçüncü bölümünde veri seti ve yöntem detaylıca açıklanmıştır. Dördüncü bölümde alınan sonuçlar raporlanmış ve son bölümde görüş ve önerilere yer verilmiştir.

2. İlgili Çalışmalar

CLEF konferanslarında düzenlenen PAN workshoplarında temel amaç yazar profili oluşturmaktır[1]. Yazar profili oluşturma görevleri yaş ve cinsiyet belirleme [3, 8], yazarın bot olup olmadığını tespiti [9], bir kullanıcının sahte haber yaymaya istekli olup olmadığı [1], metnin yanında görüntü yardımıyla cinsiyet tespiti [3] gibi alt görevlerden oluşur. Bu görevler genel olarak İngilizce, İspanyolca ve Arapça gibi dilleri kapsamaktadır. Bu görevler için veri setlerinin yanında Arapça cinsiyet tespiti [10] için detaylı veri setleri oluşturulmuş başarıları test edilmiştir.

Gutierrez ve ark. [11] İngilizce dili için farklı veri setlerini birleştirerek twitter verilerinden bot tespiti yapmaya çalışmışlardır. Yazar profili oluşturma görevlerinden olan bot tespiti otomatik hesap olarak bilinmektedir. Bu çalışma için Bert ve Roberta gibi ön eğitimli dil modellerini kullanarak 0.77 (f1-score) oranında başarımlar elde etmişlerdir.

Safara ve ark. [12] İngilizce dili için hazırladıkları çalışma da Yapay Sinir Ağları(YSA) ile birlikte Balina Optimizasyon Algoritmasını (Whale Opt. Alg. - WOA) birlikte kullanarak cinsiyet tespiti yapmaya çalışmışlardır. Elektronik posta verilerinin kullanıldığı çalışma da sınıflandırıcı olarak YSA kullanılmıştır. Ayrıca model tahminlerini iyileştirmek ve önyargıları tespit etmek için meta-sezgisel bir yöntem olan WOA kullanılmıştır. Sonuç olarak %98 doğruluk puanı elde etmişlerdir.

Zhang ve ark. [13]Arapça dili için yazar profili oluşturma ve aldatma tespiti alt görevleri için yaş, dil çeşitliliği ve cinsiyet tespiti yapmaya çalışmışlardır. Bu görev için Bert dil modelini kullanmışlardır. Farklı veri setleri üzerinde yaptıkları deneyler sonucunda cinsiyet tespiti için %81 oranında başarımlar elde etmişlerdir.

Türkçe dili için Sezerer ve ark. [4] kendi oluşturdukları Twitter veri setiyle cinsiyet tespiti yapmışlardır. Bu çalışmanın sonucunda %72.32'lik bir başarı puanı elde etmişlerdir. Ayrıca 2013 yılında Talebi ve ark. [14] tarafından yapılan çalışma da facebook üzerinden toplanan veriyi kullanarak Destek Vektör Makinaları (SVM), K-en yakın komşu (KNN) gibi klasik sınıflandırma algoritmalarının başarımlarını test etmiş cinsiyet için %90.85 yaş için %89.67 gibi başarı puanları elde etmişlerdir.

¹ <https://clef2020.clef-initiative.eu/>

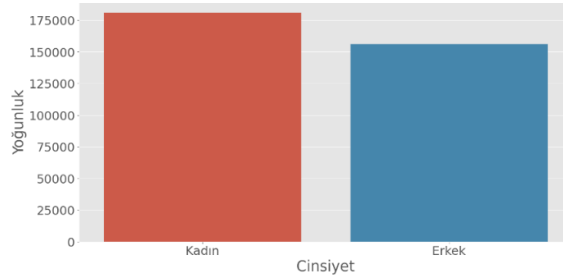
² <https://pan.webis.de/>

Cinsiyet tespiti bir metin sınıflandırma problemi gibi değerlendirilmektedir. Metin sınıflandırma görevinde önışlem adımlarından geçirilen metinler öncelikle vektör haline getirilip sonrasında sınıflandırma algoritmaları kullanılmaktadır. Önışlem adımı gereksiz kelimelerin (stop words) çıkarılması matematiksel semboller kısaltmalar gibi başarıyı düşürecek verilerin temizlenmesini kapsamaktadır[15]. Vektörleştirme işleminde ise kelime torbası(Bag of Word), N-gram terim frekansı-ters belge frekansı(TF-IDF), word2vec [16] gibi yöntemler kullanılan en ilk yöntemlerdir [15]. Sınıflandırma kısmında ise Naive Bayes, KNN [17] ve SVM gibi olasılıksal yöntemlerin yanında, tekrarlayan sinir ağı (Recurrent Neural Network-RNN) modelleri LSTM, GRU gibi derin öğrenme sınıflandırma yöntemleri kullanılmaktadır [15]. Ayrıca dikkat mekanizmasına sahip RNN ve CNN modelleri [18] metin sınıflandırmada önemli başarılar elde etmişlerdir. Dikkat mekanizması yöntemine bağlı transformer [19]tabanlı oluşturulan dil modelleri ise en son teknoloji metin sınıflandırma sonuçlarını elde etmektedirler [5, 7].

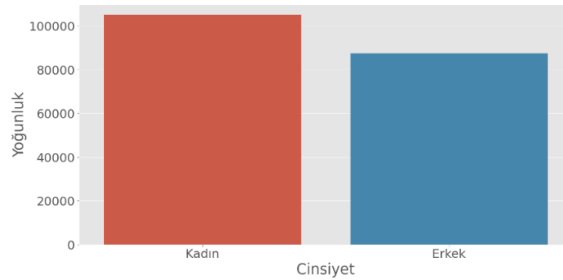
3. Materyal ve Yöntem

3.1 Veriseti

Bu çalışma için Sezerer ve ark³ [4] tarafından oluşturulan veri seti kullanılmıştır. Bu veri seti toplamda 5292 farklı kişi tarafından gönderilmiş 100'er adet gönderiden oluşmaktadır. Tüm veri seti 3368 eğitim ve 1924 test olmak üzere ayrılmıştır. Toplam gönderi sayısı ve cinsiyetlere göre dağılımı Şekil 1 ve 2'de verilmiştir. Tüm veri seti incelendiğinde kadın olarak etiketlenen veri sayısı (%53,68) kısmen fazla olsa da dengeli bir veri seti olduğu görülmektedir.



Şekil 1. Eğitim Veri Seti Cinsiyete Göre Dağılımı

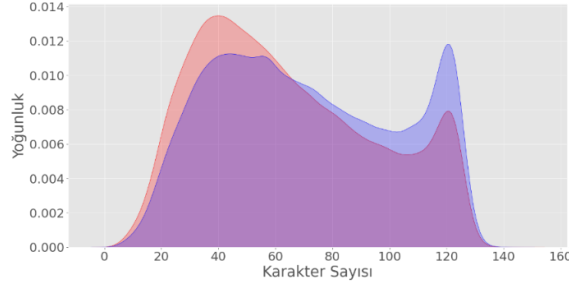


Şekil 2. Test Veri Seti Cinsiyete Göre Dağılımı

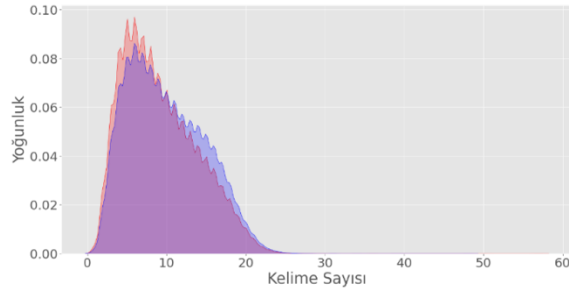
Tüm veri seti incelendiğinde (Şekil 3 ve Şekil 4) veri seti hakkında şu bilgiler gözlenmiştir;

- 25-45 karakter aralığındaki gönderilerin kadın kullanıcılar tarafından daha fazla tercih edildiği, 120 karakter gönderilerde ise durumun tam tersi olduğu,
- 40 karakterlik gönderilerde cinsiyetin kadın olma ihtimalinin erkek olma ihtimalinden daha fazla olduğu
- 120 karakterlik gönderilerde cinsiyetin erkek olma ihtimalinin kadın olma ihtimalinden daha fazla olduğu
- 110 karakter bulunduran gönderilerin her iki cinsiyet içinde beklenmeyen bir şekilde az tercih edildiği,
- 8 kelimelik gönderilerde cinsiyetin kadın olma ihtimalinin erkek olma ihtimalinden daha fazla olduğu 17 kelimelik gönderilerde durumun tam tersi olduğu,

³ <https://cloud.iyte.edu.tr/index.php/s/5DhqdlUCCdB60qG>



Şekil 3. Karakter sayısı kirmizi: kadın-mavi: erkek



Şekil 4. Kelime sayısı kirmizi: kadın- mavi: erkek

3.2. Metin ön işleme

Araştırmacıların kullanımına açık bir şekilde sunulan veri seti her bir kullanıcı için xml türü dosyalardan oluşmuştur. Bu dosyaların her birinde o kullanıcıya ait 100 adet gönderi bulunmaktadır. İlk adımda bu dosyalar sırasıyla işlenerek verilen etiketler sayesinde veri seti oluşturulmuştur. Sonrasında her bir ifade küçük harfe dönüştürülmüştür. Cinsiyet tespitinde yazılan içerikten çok yazım tarzı daha belirleyici bir özellik olduğundan “Stop Words” olarak belirtilen gereksiz kelimeler, noktalama işaretleri, emoji gibi ifadeler çıkarılmamıştır. Metinlerde bulunan sosyal medyaya özgü etiket (#), değinme (@) ve web sayfası bağlantıları (http) şeklinde etiketlenerek bu ifadelerin çeşitliliği azaltılmıştır. Bu işlemler sonucunda toplam token sayısı 420k olarak hesaplanmıştır.

3.3. Yöntem

Bu çalışma kapsamında kısa metinlerden cinsiyet tespiti yapılabilmesi için görev ikili sınıflandırma problemi olarak ele alınmıştır. Veri seti kadınlar için 0 erkekler için 1 olarak etiketlenmiştir. Eğitim aşamasında başarıyı artırabilmek için aynı yazar tarafından gönderilen metinler N boyutunda bir çerçeve kullanılarak birleştirilmiştir. $n=[1,2,4,5,10]$ olmak üzere 5 farklı değer seçilmiştir. Yazara ait yazım tarzını modelin öğrenebilmesi için çerçeve veri seti üzerinde kaydırılarak veri seti her defasında yeniden düzenlenmiştir. Tablo 1’de yeni oluşturulan veri seti ve boyutları gösterilmiştir. Her bir metin için örnek uzunluğu veri setindeki örneklerin %99’unu kapsayacak şekilde belirlenmiştir. Metin sınıflandırma algoritmaları sayesinde her bir modelin başarısı hesaplanmıştır.

Tablo 1. n çerçeve değerine bağlı olarak veri setinin yeniden oluşturulması

n	Eğitim Örnek sayısı	Test Örnek Sayısı	Örnek Uzunluğu
1	336800	192400	25
2	168400	96200	50
4	84200	48100	80
5	67360	38480	100
10	33680	19240	175

Bu çalışma kapsamında çok sayıda model oluşturularak detaylı deneyler yapılmıştır. Oluşturulan modeller 3 grup altında toplanabilir.

3.3.1. Makine öğrenmesi yöntemleri ile sınıflandırma

Klasik makine öğrenmesi modellerinde kelimeleri tokenleştirmek yani vektör haline getirmek için sınıflandırıcıdan ayrı yöntemler kullanılmaktadır. Bunlardan en çok kullanılan yöntemler N-Gram, TF-IDF ve BOW olarak bilinen kelimelerin metin içerisindeki frekanslarına ve birlikte kullanıldığı diğer kelimelere bakılarak oluşturulan yöntemlerdir [15]. Çalışmanın bu bölümünde PAN 2019 görevinde kullanılan ve yüksek başarı sağlamış yöntem oluşturduğumuz veri setleri üzerinde kullanılmıştır [20]. Tokenization işlemi için N-gram(1-3) sınıflandırma için destek vektör makineleri kullanılmıştır (N-gram + SVM).

3.3.2. Derin öğrenme yöntemleri ile sınıflandırma

Tekrarlayan sinir ağları (RNN) ve evrişimli sinir ağları (CNN) metin sınıflandırmada sıklıkla kullanılan derin öğrenme modelleri olmuşturlardır [15]. Yüksek parametre sayısı ve hesaplama gerektiren modeller olmalarına rağmen sınıflandırma problemlerinde klasik makine öğrenmesi modellerine göre başarıyı artırmışlardır[15]. Tokenleştirme işlemi BOW yöntemi kullanılarak yapay sinir ağına gömme (embedding) katmanı olarak aktarılmıştır. Vektör boyutu (embedding size) 50 olarak belirlenmiştir. Model eğitim sırasında bu vektörü devamlı güncelleyerek öğrenmeyi sağlayacaktır.

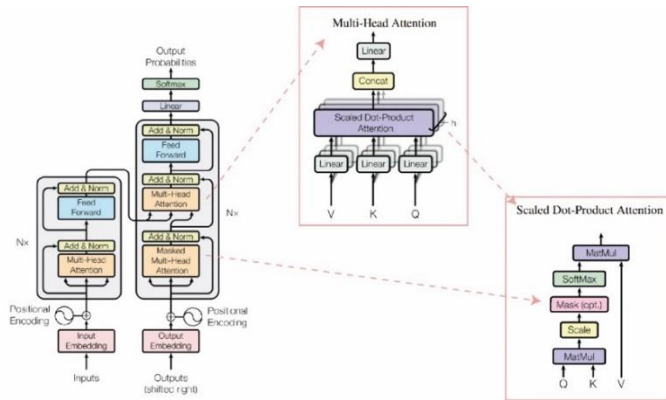
Çalışmanın bu kısmında tekrarlayan sinir ağı modeli olarak iki yönlü uzun kısa süreli bellek (Bi-LSTM)[21] modeli kullanılmıştır. 128 adet gizli katmana sahip Bi-LSTM katmanı ve çıkış katmanı olarak tek bir çıkışa sahip yapay sinir ağı eklenerek oluşturulmuştur.

Evrişimsel sinir ağı (CNN) olarak aynı şekilde 128-64-32 boyutunda 3 CNN katmanı ve tek çıkışa sahip yapay sinir ağı eklenmiştir. Her iki modelde de aşırı öğrenme problemini önlemek için dropout katmanları eklenmiş çıkış fonksiyonu sigmoid olarak seçilmiştir.

3.3.3. Ön eğitilmiş dil modelleri kullanarak sınıflandırma

Dil modelleri özetle verilen bir metinde maskelenmiş kelimeleri tahmin etmeye çalışan yapılardır. Maskelenen veri tek bir kelime olabileceği gibi bir cümle de olabilmektedir. 2017 yılında Vaswani ve ark tarafından [19] önerilen ve transformer olarak bilinen dikkat mekanizması dil modeli metin sınıflandırma, makine çevirisi ve metin özetleme gibi doğal dil işleme uygulamalarında o ana kadar ki en yüksek puanları elde etmiştir. Transformer mimarisi genel olarak Kodlayıcı (Encoder) ve kod çözücü (Decoder) olmak üzere iki ayrı ağıdan oluşur. Encoder katmanlarının çıktıları decoder katmanlarına giriş olarak verilir. Ağ n adet giriş alır ve n adet çıkış üretir. Giriş verisinden hangi veriye daha fazla dikkat edilmesi gerektiğini bulmaya çalışır.

Mimari genel olarak şu şekilde çalışır: Giriş olarak kelime embedding vektörlerini alır ve kelimelerin cümle içindeki pozisyon bilgilerini oluşturur. Bir sonraki katmanda hangi kelimeye dikkatin artması gerektiğini hesaplayabilmek için çok başlıklı dikkat mekanizmasını kullanılır. Burada girişteki her bir kelime için query, key ve value vektörleri hesaplanır. Dikkat mekanizmasından sonra giriş vektörü eklenir ve normalizasyon işlemi yapılır. Sonraki katmanda tam bağlı sinir ağı ile çıkış üretilir. Bir önceki katmanda olduğu gibi bu katmanın girişi de çıkışına eklenir ve normalizasyon işlemi yapılır(Şekil 8).

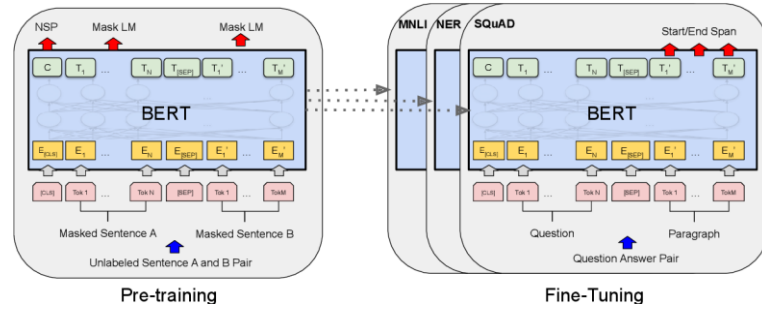


Şekil 5. Transformer ağ modeli [19]

Transformer mimarisinin başarısından sonra bu mimari kullanılarak büyük miktarda metin verileriyle (Bert:16Gb, Gpt-3:40Gb) dil modelleri oluşturulmuştur [5, 7]. Transfer öğrenmesi yöntemleri ile bu dil modelleri diğer doğal dil işleme görevlerinde (Downstream Task) kullanılmaktadır. Bu çalışma süresince Bert, DistilBert ve Electra olmak üzere 3 farklı dil modeli [22] ve versiyonları kullanılarak sınıflandırma başarıları test edilmiştir. 35Gb Türkçe metinlerle eğitilen modeller HuggingFace⁴ kütüphanesinde bulunmaktadır. Ayrıca çalışma da kelimelerin vektörle temsil edebilmek için her modelin önceden eğitilmiş token vektörleri kullanılmıştır.

3.3.3.1. Bert

Bert (Bidirectional Encoder Representations from Transformers) [5] dil modeli transformer mimarisinin encoder ağı kullanılarak wikipedia⁵ ve bookcorpus⁶ verileriyle Google beyin takımı tarafından eğitilmiştir. BertBase (Layer=12, Hidden Size=768, Self Attention Heads=12, Toplam Parametre=110M) ve BertLarge (Layer=24, Hidden Size=1024, Self Attention Heads=16, Total Parameters=340M) olmak üzere İngilizce için 2 farklı model oluşturulmuştur. Bert modelinde giriş verisinin bir kısmı mask etiketiyle maskelenir. Model eğitildikçe maskelenen kelime veya cümle verisini öğrenmeye çalışır. Böylelikle soru cevap sistemleri, varlık tanıma gibi görevlerde başarıyı artırmıştır. Ayrıca Bert modelin de giriş verisindeki kelimelerin konumlarına göre farklı vektör temsilleri bulunmaktadır. Dolayısıyla bir kelimenin cümle başında veya ortasında olmasıyla dil modelinin o kelimeye uygulayacağı dikkat değişecektir. Ön eğitimli Bert dil modelinin son katmanından sonra ek katmanlarla farklı görevler için ince ayar yapılabilmektedir. Örneğin 2'li bir sınıflandırma yapmak için tek hücreli bir sinir ağı veya soru cevap sistemi geliştirmek için cümle boyutunda bir katman eklenmesi yeterlidir (Şekil 9).



Şekil 6. Ön eğitimli bert ve Transfer öğrenme mimarisini[5]

Bu çalışmada BertBase⁷ mimarisi ile geliştirilmiş 2 farklı model kullanılmıştır. Eğitim verisi 35Gb metin dosyaları ve 44M token'dan oluşmaktadır. Kelime temsilleri birinci modelde 32k ikinci modelde 128k boyutunda vektörlerden oluşmaktadır.

3.3.3.2. DistilBert

Distilbert(Distillation Bert) damıtılmış bert olarak adlandırılır. Burada geçen damıtma terimi daha büyük bir modelin davranışını yeniden üretmek için küçük bir modelin eğitildiği sıkıştırma tekniği olarak tanımlanabilir. Parametre sayısı BertBase'in neredeyse yarısı (66M) kadardır. Bert'e göre daha kısa eğitim ve transfer öğrenme süresi sunmaktadır. Başarı olarak doğal dil anlama görevlerinde yaklaşık %2'lik bir kayıp yaşamaktadır [6].

3.3.3.3. Electra

BERT gibi maskeli dil modelleme (MLM) ön eğitim yöntemleri, bazı simgeleri [MASK] ile değiştirerek girdiyi bozar ve ardından orijinal simgeleri yeniden yapılandırmak için bir model eğitir. Doğal dil işleme görevlerine aktarıldıklarında iyi sonuçlar üretirken, genellikle etkili olmak için büyük miktarlarda hesaplama gerektirirler. Electra mimarisi alternatif olarak, değiştirilen belirteç tespiti adı verilen, örnek açısından daha verimli

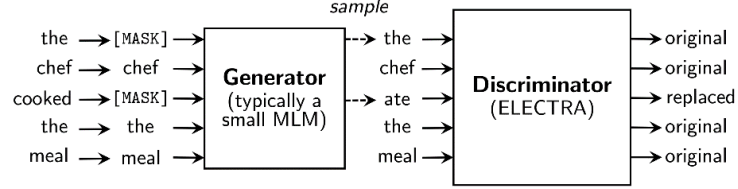
⁴ <https://huggingface.co/models>

⁵ <https://dumps.wikimedia.org/>

⁶ <https://huggingface.co/datasets/bookcorpus>

⁷ <https://huggingface.co/dbmdz>

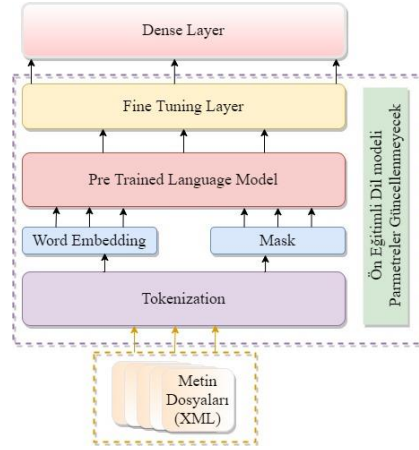
bir eğitim öncesi görev önermektedir. Bu yaklaşımda girdiyi maskelemek yerine, bazı simgeleri küçük bir jeneratör (Generator) ağından örneklenen benzer değerlerle değiştirerek onu bozar. Ardından, bozuk simgelerin orijinal değerlerini tahmin eden bir model eğitmek yerine, bozuk girdideki her bir simgenin bir jeneratör örneğiyle değiştirilip değiştirilmediğini öngören ayrımcı bir model yapısındadır [7] (şekil 10). Sonuç olarak aynı model boyutunda bert ile karşılaştırıldığında daha iyi performans göstermektedir.



Şekil 7. Electra Dil Modeli Yapısı [7]

4. Uygulama ve Sonuçlar

Bu uygulama da ön eğitilmiş dil modelleri kullanılarak Twitter gönderilerinden oluşturulmuş veri seti üzerinde cinsiyet tespiti yapılmaya çalışılmıştır. Uygulamanın genel çalışma prensibi şekil-11 de verilmiştir.



Şekil 8. Uygulama mimarisi

Çalışma kapsamında toplamda 8 adet farklı model oluşturulmuştur. Her bir model 5 farklı veri boyutuyla eğitilmiştir. Derin öğrenme yöntemleriyle ve dil modelleri kullanılarak oluşturulan modellerin parametre sayıları Tablo 2'de verilmiştir. Dil modellerinde önceden oluşturulmuş parametreler güncellenmemiş sadece son katman güncellenmiştir.

Tablo 2. Derin öğrenme modellerinin parametre sayıları

Modeller	P. Sayısı	Güncellenen P. Sayısı
Bi-LSTM		45M
CNN		50M
dbmdz/bert-base-turkish-128k-cased	185M	
dbmdz/bert-base-turkish-cased	110M	
dbmdz/distilbert-base-turkish-cased	67M	104k
dbmdz/electra-base-turkish-cased-discriminator	110M	
dbmdz/electra-base-turkish-cased-generator	35M	37k

4.1. Hiperparametreler

Eğitim süresince öğrenme oranı için 0.00001(1e-5) ile 0.005(5e-3) arası değerler test edilmiş en önemli yakınsama ‘0.001’ değerinde yakalanmıştır. Tüm modeller için aynı öğrenme oranı ve ‘adam’ optimizör kullanılmıştır. Uygulama 2’li sınıflandırma problemi olarak ele alınmış ve başarımların değerlendirilmesi için ‘binary accuracy’, kayıp fonksiyonu için ‘binary cross entropy’ seçilmiştir.

4.2. Uygulama Ortamı

Tüm uygulamalarda python programlama dili kullanılmıştır. Ön işlem adımların da ‘XML’, ‘NLTK’ kütüphaneleri makine öğrenmesi metodunda ‘SckitLearn’ kütüphanesi, Derin öğrenme ve dil modellerinin eğitimi sırasında ‘Keras’ kütüphaneleri kullanılmıştır. Ayrıca dil modellerinin önceden eğitilerek kullanılmasını sağlayan ‘transformers’ kütüphanesi kullanılmıştır. Çalışmalarda kullanılan bilgisayar 2080Ti 11Gb ekran kartı, 9. Nesil i7 işlemci ve 16 Gb ram donanımına sahiptir. Çalışmanın kodlarına ve sonuçlarına Github⁸ repomuzdan ulaşabilmektedir.

4.3. Değerlendirme

Oluşturulan modellerin başarımların ölçümleri için doğruluk (accuracy) metriği kullanılmıştır. Erkek cinsiyeti için pozitif, kadın cinsiyeti için negatif değişkenleri atanmıştır. Her bir kayıt için modelin doğru tahmin ettiği erkekler doğru pozitif (True Positive - TP), doğru tahmin ettiği kadın cinsiyeti için doğru negatif (True Negative - TN) olarak belirlenmiştir. Tam tersi olarak modelin yanlış tahmin ettiği erkek cinsiyeti için yanlış negatif (False Negative - FN), kadın cinsiyeti için yanlış pozitif (False Positive - FP) olarak belirlenmiştir. Doğruluk metriğinin hesaplanması Denklem-1 de verilmiştir [23].

$$Accuracy = (TP + TN)/(TP + FN + TN + FP) \quad (1)$$

Ayrıca en başarılı olan modelin f1-score (Denklem-2), Sensitivity (duyarlılık) (Denklem-3), Specificity (özgüllük) (Denklem-4) başarı puanları ölçülmüştür (Tablo 4).

$$F1Score = 2TP/(2TP + FP + FN) \quad (2)$$

$$Sensitivity = TP/(TP + FN) \quad (3)$$

$$Specificity = TN/(FP + TN) \quad (4)$$

4.4. Sonuçlar

Yapılan uygulamalar sonucunda ulaşılan doğruluk puanları Tablo 3’te verilmiştir. En başarılı olan modelin farklı başarımların metrikleri Tablo 4’te verilmiştir.

Tablo 3. Geliştirilen modeller ve başarı puanları

	<i>n=1</i>	<i>n=2</i>	<i>n=4</i>	<i>n=5</i>	<i>n=10</i>
TF-IDF + SVM	0.6812	0.7116	0.7540	0.7680	0.7910
Bi-LSTM	0.6823	0.7153	0.7471	0.7524	0.7660
CNN	0.6781	0.6830	0.6981	0.7158	0.7459
Bert	0.6967	0.7163	0.7524	0.7651	0.7931
Bert-128k	0.7017	0.7304	0.7630	0.7708	0.8012
Distilbert	0.6452	0.6525	0.6824	0.7028	0.7457
Electra Discriminator	0.6825	0.7157	0.7383	0.7451	0.7747
Electra Generator	0.6910	0.7238	0.7538	0.7657	0.7952

⁸ https://github.com/ilhamisel/Gender_Identification

Tablo 4. Bert-128k modelinin farklı başarımları metrikleri

	<i>TP</i>	<i>TN</i>	<i>FN</i>	<i>FP</i>	<i>Doğruluk</i>	<i>F1-Score</i>	<i>Duyarlılık</i>	<i>Özgüllük</i>
<i>n=1</i>	61198	73809	26202	31191	0.7017	0.6808	0.7002	0.7029
<i>n=2</i>	31868	38396	11832	14104	0.7304	0.7108	0.7292	0.7314
<i>n=4</i>	16642	20059	5208	6191	0.7630	0.7449	0.7616	0.7642
<i>n=5</i>	13483	16177	3997	4823	0.7708	0.7535	0.7713	0.7703
<i>n=10</i>	7002	8413	1738	2087	0.8012	0.7855	0.8011	0.8012

5. Tartışma

Bu çalışma kapsamında Türkçe Twitter gönderilerinden cinsiyet tespiti yapılmıştır. Problem, sınıflandırma görevi gibi ele alınıp klasik makine öğrenmesi yöntemleri, derin öğrenme modelleri ve ön eğitilmiş dil modelleri kullanılmıştır. Yapılan deneyler sonucunda en başarılı model bert mimarisi ile hazırlanan ve kelime boyutunu 128k vektörle temsil eden model olmuştur. Derin öğrenme mimarileri ile karşılaştırıldığında çok daha az parametre eğitimi ile çok daha yüksek puan almıştır. Electra Generator mimarisine sahip dil modeli ise en az parametreye sahip olmasına rağmen yüksek oranda başarıyı yakalamıştır. Dolayısı ile bu çalışma da Türkçe dil modellerinin az sayıda parametre eğitimi ile sınırlı sayıda eğitim seti ve kısa ve düzensiz kelime grupları üzerinde doğal dili anlama becerileri ölçülmüştür. Önerilen model ve yöntemlerin farklı metin sınıflandırma problemlerinde kullanılabilir olduğu gösterilmeye çalışılmıştır.

Veri seti içeriği düzensiz gönderiler, kısa metinler, farklı yazım yanlışları, çok sayıda kelime türü gibi sınıflandırmayı etkileyen olumsuz öğelerden oluşmaktadır. Başarımı artırmak için farklı yöntemlere ihtiyaç duyulmaktadır. Bu sorunu giderebilmek için bu çalışmada gönderileri tek başına değerlendirmenin yanında belirlediğimiz bir çerçeveye boyutuyla ek çalışmalar yapılmıştır. Sonuç olarak tek bir tweet için cinsiyet tespiti zor olmasına rağmen 5-10 arası tweet gönderileri değerlendirmeye alındığında %80'e kadar doğru tahmin edilebilmektedir. Çalışma Türkçe gönderilerden cinsiyet tespiti için hazırlanan en kapsamlı çalışmalardan biri olmaktadır. Ayrıca kısa metin sınıflandırma görevi olarak ele alındığında ise ön eğitilmiş dil modellerinin Türkçe için başarımları kıyaslanarak sonraki çalışmalar için bir rehber olmaktadır.

Aynı veri seti ile ulaşabildiğimiz tek yayın olan diğer çalışmada [4] (Bag-Of-Words+SVM) %72.32 oranında başarıya ulaşıldığı belirtilmektedir. Yazarların veri setini ham bir şekilde paylaştığından ve metin ön işleme adımları detaylıca belirtilmediğinden dolayı bu çalışma da kullanılan SVM sınıflandırma yöntemi ile aynı sonuçlara ulaşamamıştır. Oluşturulan en başarılı Bert modeli ile $n=1$ için %70 $n=2$ için %73 başarımları elde edilmiştir.

Ayrıca ön eğitilmiş dil modellerinin eğitimi için yüksek miktarda veri kullanılmaktadır. Bu veriler çoğunlukla düzenli, anlamlı ve standart cümle yapılarından oluşmaktadır. Dil modellerinde kelimelerin cümle içerisindeki anlamı, diğer kelimelerle ilişkisi hatta konumu bile kodlanmaktadır (encoding). Bu yüzden metin sınıflandırma, duygu analizi gibi uygulamalarda klasik makine öğrenme modellerine oranla yüksek başarımlar elde edebilmektedirler. Fakat twitter gibi düzensiz kelime öbeklerinden oluşan, sohbet dili kullanılmış metin kümeleri üzerindeki uygulamalarda aynı oranda fark oluşmamaktadır. SVM gibi sınıflandırıcılar dil modellerinin aksine metin içerisindeki cümlelerin yapısıyla veya kelimelerin anlamları ile ilgilenmezler. Bu sebeple klasik makine öğrenme algoritmaları cinsiyet belirleme görevi için yüksek parametrelili dil modellerine yakın sonuçlar elde etmiştir.

Kaynaklar

- [1] F. M. R. Pardo, A. Giachanou, B. Ghanem, and P. Rosso, "Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter," *CLEF 2020 Labs Work. Noteb. Pap.*, pp. 22–25, 2020.
- [2] M. A. Álvarez-Carmona *et al.*, "A visual approach for age and gender identification on Twitter," *J. Intell. Fuzzy Syst.*, vol. 34, no. 5, pp. 3133–3145, 2018, doi: 10.3233/JIFS-169497.
- [3] F. Rangel, P. Rosso, M. Montes-Y-Gómez, M. Potthast, and B. Stein, "Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter," *CEUR Workshop Proc.*, vol. 2380, 2018.
- [4] E. Sezerer, O. Polatbilek, and S. Tekir, "A Turkish Dataset for Gender Identification of Twitter Users," pp. 203–207, 2019, doi: 10.18653/v1/w19-4023.
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," no. M1m, 2018, [Online]. Available: <http://arxiv.org/abs/1810.04805>.
- [6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv*, pp. 2–6, 2019.

- [7] C. D. Manning, “Electra : P Re - Training T Ext E Ncoders As D Iscriminators R Ather T Han G Enerators,” *Iclr*, pp. 1–18, 2020.
- [8] F. Rangel, P. Rosso, M. Potthast, and B. Stein, “Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter,” *CEUR Workshop Proc.*, vol. 1866, 2017.
- [9] F. Rangel and P. Rosso, “Overview of the 7th author profiling task at Pan 2019: Bots and gender profiling in twitter,” *CEUR Workshop Proc.*, vol. 2380, 2019.
- [10] W. Zaghouani and A. Charfi, “Arap-Tweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification,” *arXiv*, pp. 694–700, 2018.
- [11] D. Martin-Gutierrez, G. Hernandez-Penalosa, A. B. Hernandez, A. Lozano-Diez, and F. Alvarez, “A Deep Learning Approach for Robust Detection of Bots in Twitter Using Transformers,” *IEEE Access*, vol. 9, pp. 54591–54601, 2021, doi: 10.1109/ACCESS.2021.3068659.
- [12] F. Safara *et al.*, “An Author Gender Detection Method Using Whale Optimization Algorithm and Artificial Neural Network,” *IEEE Access*, vol. 8, pp. 48428–48437, 2020, doi: 10.1109/ACCESS.2020.2973509.
- [13] C. Zhang and M. Abdul-Mageed, “BERT-based Arabic social media author profiling,” *CEUR Workshop Proc.*, vol. 2517, no. 1, pp. 84–91, 2019.
- [14] M. Talebi and C. Köse, “Facebook yorumlarının analiziyle Cinsiyet, Yaş ve Eğitim düzeyi belirleme Identifying Gender, Age and Education level by analyzing comments on Facebook,” *Ieee*, no. 2007, pp. 4–7, 2013.
- [15] L. I. Qian *et al.*, “A Survey on Text Classification: From Shallow to Deep Learning,” *arXiv*, vol. 31, no. 11, pp. 1–21, 2020.
- [16] İ. Sel, A. Karci, and D. Hanbay, “Karşılıklı Bilgi Kullanılarak Metin Sınıflandırma İçin Özellik Seçimi Feature Selection for Text Classification Using Mutual Information,” *2019 Int. Artif. Intell. Data Process. Symp.*, pp. 18–21, 2019.
- [17] İ. Sel and D. Hanbay, “Doğal Dil İşleme Yöntemleri Kullanarak E- Maillerin Sınıflandırılması E- Mail Classification Using Natural Language Processing,” *2019 27th Signal Process. Commun. Appl. Conf.*, pp. 19–22, 2019.
- [18] D. Bahdanau, K. H. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.
- [19] A. Vaswani *et al.*, “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.
- [20] J. Pizarro, “Using N-grams to detect Bots on Twitter Notebook for PAN at CLEF 2019,” *CEUR Workshop Proc.*, vol. 2380, pp. 9–12, 2019.
- [21] S. ElSayed and M. Farouk, “Gender identification for Egyptian Arabic dialect in twitter using deep learning models,” *Egypt. Informatics J.*, vol. 21, no. 3, pp. 159–167, 2020, doi: 10.1016/j.eij.2020.04.001.
- [22] S. Schweter, “BERTurk - BERT models for Turkish.” Zenodo, 2020, doi: 10.5281/zenodo.3770924.
- [23] Y. Altuntaş and K. Fatih, “Deep Feature Extraction for Detection of Tomato Plant Diseases and Pests based on Leaf Images,” *Celal Bayar Üniversitesi Fen Bilim. Derg.*, vol. 17, no. 2, pp. 145–152, 2021, doi: 10.18466/cbayarfbe.812375.