Türkiye Elektrik Mühendisliği ve Bilgisayar Dergisi
Bilimler

Cilt 31 Sayı 3 Madde 7

5-1-2023

Çok sınıflı yanlış pozitif azaltma modeliyle derin öğrenme tabanlı Türkçe yazım hatası tespiti

BURAK AYTAN

CEMAL OKANSAKAR

Türkiye Elektrik Mühendisliği ve Bilgisayar Bilimleri Dergisi

http : // j tablolarımız . tamam tamam. hükümet tr / el ek trik/

Araştırma Makalesi

Turk J Elec Eng & Comp Sci
(2023) 31: 581 – 595

doi:10.55730/1300-0632.4003

Çok sınıflı yanlış algılama ile derin öğrenme tabanlı Türkçe yazım hatası tespiti pozitif indirgeme modeli

Burak AYTAN[1,*], **C.Okan SAKAR**[2]

[1]Bilgisayar Mühendisliği Bölümü, Bahçeşehir Üniversitesi, İstanbul, Türkiye,
[2]Mühendislik ve Teknoloji Fakültesi, Orta Doğu Amerikan Üniversitesi, Kuveyt

Kabul edilmiş:28.09.2022 • **Kabul Edildi/Çevrimiçi Yayınlandı:**15.03.2023 • **Son sürüm:**28.05.2023

Soyut: Yazım denetimi ve düzeltme, metin normalleştirme sürecinde önemli bir adımdır. Türkçe gibi eklemeli dillerde bu görevler daha zordur çünkü pek çok ek bir araya getirilerek birçok kelime kök kelimeden türetilebilir. Bu çalışmada Türkçede yanlış yazılan kelime tespiti için iki aşamalı derin öğrenme tabanlı bir model öneriyoruz. İnternet paylaşım platformlarında yaygın olarak kullanılan yabancı kelime ve kısaltmaların kullanımından kaynaklanan hatalı pozitif tahminleri azaltmak için sisteme yanlış pozitif azaltma modeli entegre edilmiştir. Bu amaçla etiketlemeye yönelik bir mobil uygulama geliştirerek çok sınıflı bir veri seti oluşturuyoruz. Uzun Kısa Süreli Bellek (LSTM) ve Çift Yönlü LSTM (Bi-LSTM) ağları ile birlikte karakter tabanlı, hece tabanlı ve bayt çifti kodlama (BPE) yaklaşımları dahil olmak üzere farklı simgeleştirici türlerinin kullanılmasının etkisini karşılaştırıyoruz. Bulgular, önerilen BPE tokenizerli Bi-LSTM tabanlı modelin kıyaslama yöntemlerinden üstün olduğunu gösteriyor. Sonuçlar aynı zamanda yanlış pozitif azaltma adımının, geri çağırmada nispeten daha az bir düşüş karşılığında baz tespit modelinin hassasiyetini önemli ölçüde arttırdığını göstermektedir.

Anahtar kelimeler: Metin normalleştirme, yazım denetleyici, belirteçler, uzun kısa süreli bellek, eklemeli diller

1. Giriş

İnternet kullanımının artmasıyla birlikte güvenilir doğal dil işleme (NLP) araçlarının oluşturulmasına olan ihtiyaç da artıyor. NLP görevleri, kullanıcılar tarafından üretilen metni Chatbot geliştirme gibi farklı amaçlar için girdi olarak kullanır.1], soru yanıtlama [2], otomatik konuşma tanıma [3], konuşma işleme [4], duygu analizi [5,6], metin sınıflandırması [7] ve makine çevirisi [8] Şirketler müşterilerine daha iyi hizmet verebilmek ve performanslarını artırmak için bu NLP görevlerine dayalı yazılım ve uygulamalar geliştirmektedir.
Pazar payı.

İnternet topluluğu tarafından üretilen devasa miktardaki veri,
son derece doğru NLP araçları oluşturmak. Ancak internet platformlarından alınan kullanıcı metninin daha az standart bir dille yazılması ve yüksek oranda yanlış yazılmış kelimeler içermesi, ilgili NLP görevleri için geliştirilen modellerin performansını kötüleştirmektedir.9–11] Bu nedenle metin normalleştirme, daha iyi modeller elde etmek için önemli bir adımdır ve aynı zamanda anadili dil olmayan okuyucuların içeriği daha iyi anlamasını sağlar.

Yazım hatalarının tespiti ve düzeltilmesi, metin normalleştirmesi için kritik bir adımdır. Bu çalışmamızda Türkçedeki yazım hatası tespit problemleri üzerinde durulmaktadır. İngilizce, Fransızca ve Almanca gibi çoğu dil için yazım denetimi, kural tabanlı veya sözlük tabanlı yaklaşımlarla ele alınabilir; ancak

*Yazışma adresi: burak.aytan@bahcesehir.edu.tr

Türkçe, Fince ve Macarca gibi kelimelerin anlamının kelime kökü ve biçimbirimin birleşimiyle belirlendiği eklemeli dillerde yazım denetimi de morfolojik analiz gerektirir.12–15] Bu tür dillerde eklerin köke eklenmesiyle sözcüğün anlamı değişir.

Yazım hataları genel olarak sözcük dışı hatalar (NWE) ve gerçek sözcük hataları (RWE) olarak iki kategoriye ayrılır.16,17] Yanlış yazılan metin dilde var olan bir kelimeyse buna gerçek kelime hatası (RWE), aksi halde kelime dışı hata (NWE) adı verilir. Örneğin, İngilizce'deki 'fun' kelimesi 'gun' veya 'vun' olarak yanlış yazılmış olabilir; bu durumda ilki RWE, ikincisi ise NWE'dir [16] Bu tür hatalar İngilizce'de daha yaygındır ve anlam ve bağlam analizi gerektirdiğinden bu tür hataları tespit etmek zordur. Çalışmamızda Türkçede en sık rastlanan hata türü olduğunu düşünerek NWE'nin tespitine odaklandık.

Bu çalışmanın motivasyonu, Türkçe için yüksek doğruluklu bir yazım hatası tespit modeli önermektir. Sondan eklemeli dillerde çok sayıda ek kullanılması nedeniyle, önceden belirlenmiş kuralların kullanımına dayalı yöntemler, ilgili NLP görevlerinde iyi performans göstermemektedir. Bu modeller aynı zamanda internet paylaşım platformlarında yaygın olarak kullanılan yabancı kelime ve kısaltmalar için de hatalı pozitif tahminler üretmektedir. Bu çalışmanın katkıları şu şekildedir:

- Yüksek tespit oranına sahip yazım tespiti için derin öğrenmeye dayalı bir model oluşturmak.
- Türk web camiasında yaygın olarak kullanılan yabancı dil ve kısaltmaları içeren etiketli veri seti oluşturmaya yönelik mobil uygulama geliştirilmesi.
- Yabancı kelimelerin, kısaltmaların veya temel tespit modelinin tespit edemediği durumların kullanımından kaynaklanan hatalı pozitif tespitlerin sayısını azaltmak için hatalı pozitif azaltma modelinin entegre edilmesi.
- Türkçe dili için yazım hatası tespit problemlerinde farklı tip belirteçlerin kullanılmasının etkisinin karşılaştırılması.

Bu çalışmanın geri kalanı şu şekilde organize edilmiştir. Kısımda2yazım hatalarının tespitine odaklanan çalışmalara genel bir bakış sunulmaktadır. Bölüm3Bu çalışmada kullanılan veri kümelerinin bir açıklamasını sağlar. Önerilen derin öğrenmeye dayalı tespit modellerinin detayları Bölüm'de sunulmaktadır. 4. Bölüm5Deneysel sonuçları ve ilgili tartışmayı verir. Bölüm6çalışmayı özetler ve sonuçlandırır.

2. İlgili çalışma

Yanlış yazılmış kelime tespit görevine yönelik uygulamaların uygulanması tamamen kelime sözlükleri veya ağaç yapıları gibi dış kaynaklara dayanır. Genel olarak bu yöntemler, İngilizce gibi daha az karmaşık kelime yapılarına sahip diller için iyi çalışır.18] Ancak Türkçe gibi morfolojik karmaşıklığı yüksek olan eklemeli diller için kapsamlı kaynaklar oluşturmak zordur. İlgili çalışmalarda önceden tanımlanmış kurallara, en kısa mesafe algoritmalarına, metin benzerliği metriklerine ve derin öğrenmeye dayalı farklı yöntemler kullanılmıştır. Bu bölümde Türkçede yazım hatası tespitine odaklanan çalışmalara genel bir bakış sunuyoruz.

Solak ve Oflazer [13] üç adımdan oluşan kural tabanlı bir yöntem önerdi. Bu adımlar kök belirleme, morfofonemik kontroller ve morfolojik ayrıştırmadır. Ana fikir, ilk önce maksimum eşleşme algoritmasına dayalı olarak verilen metnin kökünü bulmaktır. Bu algoritmada öncelikle kelimenin tamamı aranır.

sözlükte yalnızca kökleri ve bazı düzensiz gövdeleri içeren. Bulunursa, kelimenin hiçbir eki olmadığı ve bu nedenle ayrıştırılmasına gerek olmadığı varsayılır. Aksi takdirde, kelimenin en sağındaki bir harf çıkarılır ve ortaya çıkan alt dize, kelimenin tamamıyla tekrar karşılaştırılır. Bu işlem kök bulununcaya kadar devam eder. Kelimenin kökü bulunduktan sonra kalan kısım ek olarak değerlendirilerek morfolojik açıdan incelenir. Morfolojik analiz yapılırken Türkçenin dilbilgisi kuralları belirlenmekte ve manuel olarak tanımlanan algoritmalar uygulanmaktadır. Morfolojik analize yönelik bir diğer yaklaşım ise kelimenin kökü bulunduktan sonra uygulanan dinamik programlama tabanlı arama algoritmalarıdır.19]

Akın ve Akın [20] ayrıca sözlük kullanan kural tabanlı bir yöntem önerdi. Sözlük, kelimelerin köklerinden ve bunların çeşitli biçimlerinden oluşur. Kelimenin kökü bulunurken öncelikle aday kök bulunur ve daha sonra Direkt Asiklik Kelime Grafiği (DAWG) algoritmasının uygulanmasıyla arama işlemi devam eder. 21] Morfolojik ayrıştırıcı bölümünde algoritma, ek alternatif kalmayıncaya kadar köke olası son ekleri eklemeye devam eder. Önerilen yazım denetleyicisi morfolojik bir ayrıştırıcı ve Damerau-Levenshtein Düzenleme Uzaklığı Algoritmasını kullanır [22] Özel isim tespiti ve ünlü restorasyon işlemleri ile yedi normalleştirme katmanından oluşan kural tabanlı bir başka yöntem Eryiğit ve Torunoğlu tarafından önerilmiştir.23] Bu yöntem, yanlış yazılan kelime için tek bir aday kelime üretir. Her katman, sosyal medya metinlerinde görünen belirli hata türlerini ele alacak şekilde tasarlanmıştır. Bu yöntemde TDK'dan (Türk Dil Kurumu) alınan kısaltma listesi ve bazı yabancı kelimeler de dikkate alınmıştır.

Daha yeni çalışmalardan birinde Çolakoğlu ve ark. [24] Türkçe metnin normalleştirilmesi için bir makine çevirisi yaklaşımı tasarladı. Bu çalışmada hem istatistiksel makine çevirisi hem de sinirsel makine çevirisi yöntemlerini kullandılar. 6-g karakter düzeyinde bir dil modeli olan KenLM [25], istatistiksel makine çevirisi bölümünde ve bir kodlayıcı-kod çözücü mimarisi olan OpenNMT kullanıldı [26], sinirsel makine çevirisi kısmında kullanıldı. Başka bir çalışmada Büyük [27] Türkiye normalizasyonu için bir diziden diziye model önerdi. Seq2seq modeli LSTM tabanlı bir modeldi ve sözcükleri harflere ayırarak karakter tabanlı bir tokenizer kullanıldı. Bu çalışmanın diğer seq2seq modelleriyle karşılaştırıldığında katkısı esas olarak tokenizasyon kısmında oldu. Bu görev için, yazım dışı sözcüklerin ilk üç ünsüz harfi seq2seq modeline girdi olarak beslendi.

Safaya ve ark. [28] Hunspell tabanlı bir [29] Türk dili için yöntem. Hunspell, başlangıçta Macar dili için tasarlanmış bir tür yazım denetleyicisidir. Zengin morfoloji, karmaşık kelime kombinasyonları ve karakter kodlamasıyla birçok dile uygulanmıştır.29] Yakın zamanda yapılan başka bir çalışmada [30], hatalı kelimeleri tespit etmek için bir morfolojik analizör kullanıldı ve bu analizör tarafından bir morfem dizisine ayrıştırılamayan kelimeler standart olmayan olarak tanımlandı. Daha sonra standart dışı olarak tanımlanan kelimelerin tamamı Türkçe varlık tanıyıcıdan geçirilerek tanınan kelimeler elenmiştir.

3. Veri kümesi açıklaması

Bu bölümde deneylerde kullanılan eğitim ve test veri setlerinin açıklamasını sunuyoruz. Ayrıca hatalı yazılan kelimelerin oluşturulmasında kullanılan işlevler hakkında da detaylı bilgi veriyoruz.

3.1. Eğitim veri seti

Çalışmamızda tespit modellerinin eğitimi ve test edilmesi için farklı kaynaklardan çeşitli veri setleri kullanıldı. Doğru yazılan kelimelerin yer aldığı veri seti, TDK'dan (Türk Dil Kurumu) alınan sözlük kullanılarak oluşturulmuştur. Bu veri setinin kişi ve zaman eklerini içeren genişletilmiş versiyonu da oluşturuldu. Daha sonra bu verilere kişi ve yer adlarını içeren özel isimler eklenerek özetleme yapılmıştır.

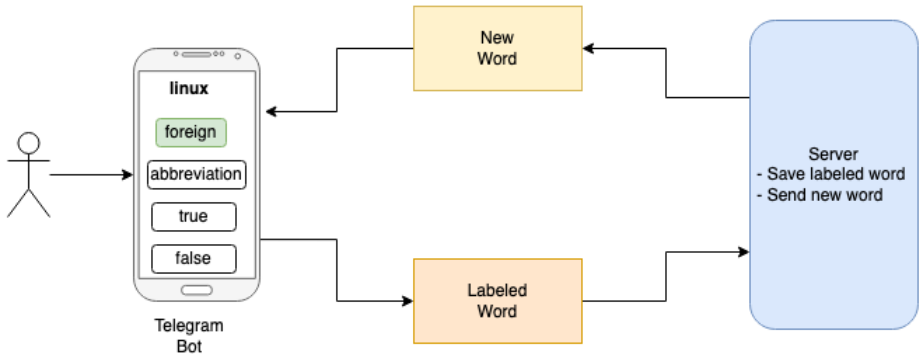
1 milyon kelime.

TDK'dan elde edilen veri seti yazım hatası olmayan kelimelerden oluşmaktadır. Bu nedenle doğru kelimelerden yanlış yazılan kelimeler oluşturmak için bazı işlemler uyguladık. Bu amaç için kullanılan işlevler Tabloda gösterilmektedir. 1. Sonuç olarak 1 milyon doğru ve 7 milyon yanlış yazılan kelimeden oluşan bir veri seti elde edildi. Ana tespit modeli bu veri seti kullanılarak oluşturuldu.

tablo 1. Model eğitimi için yanlış yazılan sözcüklerin oluşturulmasına yönelik geliştirilen işlevler.

Fonksiyon adı	Detay	Örnek
Sesli harfi kaldır	Kelimelerdeki sesli harflerin bir kısmı veya tamamı silinir	“merhaba”→“merhba”
Sesli harf ekle	Kelimedeki bazı sesli harfler kopyalanıyor	“geliyorum”→“geeliyorum”
Asciify	Türkçe karakterler İngilizce karakterlere dönüştürülür	“ağlıyorum”→“aglıyorum”
İki karakteri değiştirme	Ardışık iki karakterin yerini rastgele değiştirme	“kelime”→“keilme”
Son karakter ekleniyor	Kelimenin son karakterini çoğaltma	“erzurum”→“erzurumm”
Bilinen hatalar	Türkçede en sık karşılaşılan yazım hataları	”seviyorum”→“seviyom”

Yanlış yazılan kelimeler olarak sınıflandırılan örnekler çalışmamızda olumlu tahminler olarak değerlendirilmektedir. Yanlış pozitif azaltma modeli, yanlış yazılmış olarak etiketlenen doğru yazılmış sözcükler olan temel modellerin yanlış pozitif tahminlerini daha da azaltmak için tasarlanmıştır. Bu modeli eğitmek için öncelikle Türkçe Vikipedi, Türkçe OSCAR ve bazı haber sitelerinden elde edilen 38 GB büyüklüğündeki verilere taban tespit modeli uygulanmıştır. Bu adımda yanlış yazılan (olumlu tahminler) olarak sınıflandırılan kelimeler, görülme sıklıklarına göre büyükten küçüğe doğru sıralanmıştır. Daha sonra Şekilde gösterilen tespit modelinin tahminlerinin onaylanması ve düzeltilmesi amacıyla en sık kullanılan kelimelerin gözden geçirilmesini sağlayan bir mobil uygulama geliştirilmiştir.1. Bu listedeki kelimeler bu etiketleme işlemi sırasında dört kategori altında kategorize edilmiştir: (1) doğru şekilde pozitif olarak sınıflandırılmış (yanlış yazılmış olduğu tahmin edilen kelimeler), (2) yanlış olarak pozitif olarak sınıflandırılmış (yanlış yazıldığı tahmin edilen, doğru yazılmış kelimeler)), (3) kısaltma ve (4) yabancı kelime. Bu aşamada incelenen kelime sayısı 10.000 olup bunların %35'i hatalı, %30'u yabancı, %20'si kısaltma ve %15'i doğru yazılmıştır.



Şekil 1. Sistemin algılama yeteneğinin arttırılması amacıyla kelime bazlı etiketlemeye yönelik geliştirilen mobil uygulama.

3.2. Veri kümesini test edin

Test setinin orijinal setten çıkarıldığı klasik bir çapraz doğrulama prosedürünün kullanılması, yanlış yazılan kelimeleri bazı önceden tanımlanmış işlevlerle yapay bir şekilde oluşturduğumuz için çalışmamızda taraflı sonuçlara yol açabilir. Bu nedenle mikroblog sitesi Twitter'dan alınan kelimeleri manuel olarak etiketledik ve kullanıcı metinlerinden oluşan test verileri oluşturduk. Ulaşılan içerikten 1974'ü doğru, 448'i hatalı olmak üzere 2422 kelime elde ettik. Eğitim seti kullanılarak yanlış pozitif azaltma modeli olan ve olmayan tespit sistemi oluşturulmuş ve elde edilen model, gerçek kullanıcı metninden oluşan bu test setine uygulanmıştır.

4. Önerilen yöntemler

Bu çalışmada iki yazım hatası tespit modeli öneriyoruz. İlk model, her kelimeyi doğru yazılmış veya yanlış yazılmış olarak sınıflandıran temel tespit modelidir. İkinci model, temel tespit modeli tarafından yanlış yazılmış olarak sınıflandırılan kelimelere uygulanan başka bir yanlış pozitif azaltma modelini içerir. Yanlış pozitif indirgeme modeli, her kelimeyi doğru pozitif, yanlış pozitif, yabancı ve kısaltma sınıflarından birine atayan çok sınıflı bir sınıflandırma problemi olarak tasarlanmıştır. Bu bölümde öncelikle her iki tespit modeli için kullanılan tokenizerlere genel bir bakış sunuyoruz. Daha sonra temel tespit modelinin, yanlış pozitif indirgeme modelinin ve önerilen birleşik tespit modelinin ayrıntılarını veriyoruz.

4.1. Tokenlaştırıcılar

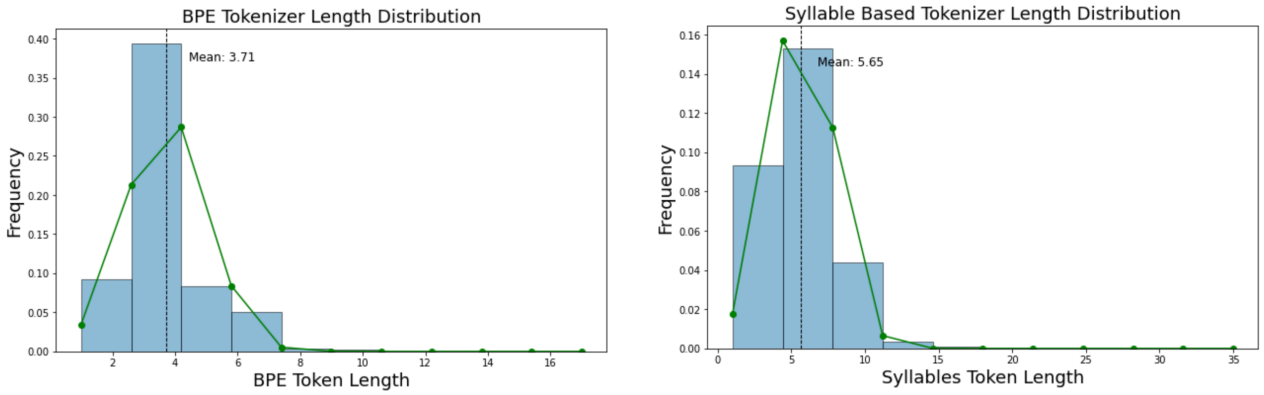
Tokenizasyon, bir cümleyi, paragrafı veya metni kelimelere veya daha küçük parçalara ayırma işlemidir.31] Bu, özellikle eklemeli diller için son derece doğru bir yazım hatası tespit modelinin oluşturulmasında çok önemli bir adımdır. İşlemin sonunda oluşan her parçaya token adı verilir [31] Çalışmamızda uyguladığımız ayrıntıları Tabloda görülebilen üç tip tokenizer2. Karakter bazlı bir belirteç kullanmak, sözcüklerin tek tek karakterlere bölünmesini içerir; böylece her karakter bir belirteç oluşturur. Alternatif olarak, ikinci yaklaşımımız, Tabloda sunulan bilgilerde örneklendiği gibi kelimelerin hecelere bölündüğü ve her hecenin daha sonra bir simge olarak kullanıldığı hece bazlı bir belirteç ayırıcının kullanımını içeriyordu.2.

Tablo 2. Tespit modellerinde tokenizasyon işlemi için kullanılan tokenlaştırıcılara genel bakış.

Tokenlaştırıcı	Detay	Örnek
Karakter tabanlı belirteç	kelimenin her karakteri bir belirteç olarak kullanılır	“istanbul”→ ['İstanbul']
Hece tabanlı belirteç	Kelimeler hecelere bölünür ve her hece bir simge olarak kullanılır.	“konuşuyorum”→ ['ko','nu','şu','yo','rom']
Bayt Çifti kodlama (BPE) belirteci	Bu belirteç, külliyattaki karakter gruplarının sıklığına bakarak belirteçler oluşturur	“sabah edebilmek”→ ['sabah','yapabilme','k']

Karakter ve hece tabanlı belirteçlerle birlikte bir bayt çifti kodlayıcı (BPE) belirteç oluşturucu ekledik [32] BERT (Transformatörlerden Çift Yönlü Kodlayıcı Gösterimleri) gibi dil modellerinde yaygın olarak kullanılan bir teknik olan metodolojimize dahil edilmiştir. Bu yaklaşım, Tablo 1'de ana hatlarıyla belirtildiği gibi, bütünlük içinde mevcut karakter gruplamalarının sıklığını değerlendirerek belirteçlerin oluşturulmasını içerir.2. BPE belirteçleyicinin uygulanmasıyla, sık kullanılan kelimeler belirteçler olarak kategorize edilir ve tüm son ekler

daha sonra ayrı jetonlar olarak işlenebilir. BPE tokenizer, tokenları, corpus içinde gözlemlenen karakter gruplamalarının sıklığına göre belirlediğinden, korpusun rolü bu yöntemin başarısı için temeldir. Buna göre, iki farklı derlem kullanarak BPE tokenizerini kullandık. İlk yineleme yalnızca doğru yazılan sözcükler kullanılarak oluşturuldu, ikinci yineleme ise hem doğru hem de yanlış yazılan sözcükleri içeren daha kapsamlı bir derleme kullanılarak formüle edildi. Her iki yinelemeden elde edilen sonuçlar, sonuçlar bölümünde ayrıntılı olarak açıklanmıştır. Figür2hem BPE hem de hece tabanlı belirteç oluşturucularla elde edilen belirteç uzunluklarının dağılımını gösterir.

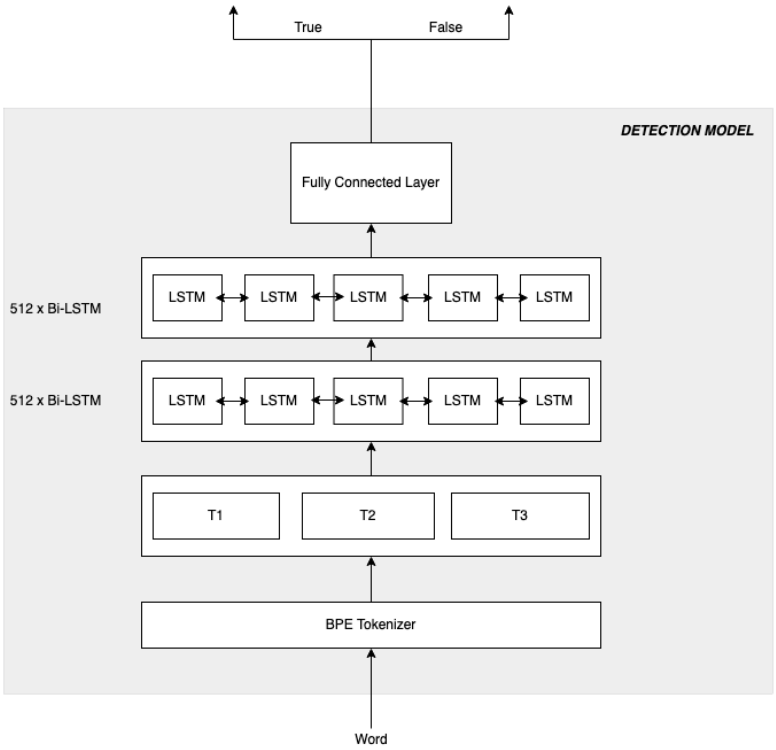


şekil 2. Tokenizer tabanlı token uzunluğu dağılımı.

4.2. Baz algılama modeli

Şekilde gösterilen temel algılama modeli3iki gizli Bi-LSTM katmanından oluşan ikili sınıflandırma problemi olarak tasarlanmıştır. Bu model girdi olarak bir kelimeyi alır ve kelimenin doğru yazılıp yazılmadığını çıktı olarak verir. LSTM mimarisinin tekrarlayan bağlantıları, yazım hatalarının tespitinde modelin önceki karakter veya kelime parçalarını sonrakilerle birlikte değerlendirmesini sağlar. LSTM ve transformatör mimarileri arasındaki en önemli fark, uzun menzilli bağımlılıkları ele alma biçimleridir. Transformatör mimarisi, uzun menzilli bağımlılıkları açıkça modellemek için öz-dikkat mekanizmalarını kullanır ve bu da onu LSTM'ye kıyasla uzun dizileri işlemeye daha uygun hale getirir.33,34] Ancak önerilen yazım hatası tespit modeli kelime tabanlı olduğundan girdisi kapsamlı dizilerin işlenmesini gerektirmez. Özellikle Şekilde gösterildiği gibi2, BPE ve hece tabanlı belirteçlerle ilişkili ortalama belirteç uzunluğu:3.71 Ve5.65,sırasıyla belirteç uzunluklarının %99'u BPE için 9'dan az ve hece tabanlı belirteç oluşturucu için 15'ten azdır. Bu nedenle önerilen tespit modelinde transformatör tabanlı model yerine LSTM tabanlı modeli seçiyoruz.

Şekilde görüldüğü gibi3Önerilen modelin mimarisi, her birinde 512 LSTM düğümü bulunan iki gizli katmandan oluşmaktadır. Model, kelimenin tamamını girdi olarak alır ve daha sonra Tabloda verilen belirteçleri kullanarak onu belirteçlere ayırır.2. Bu işlemi 300 düğüm içeren bir gömme katmanı takip etmektedir. Elde edilen temsiller, her biri 512 Bi-LSTM düğümünden oluşan iki gizli katmana beslenir. Gizli katman temsillerini ikili çıkışlara eşleyen tam bağlantılı katman 1024 düğümden oluşur. Bu model farklı tokenizerler, tokenizasyon için kullanılan korpus ve Bi-LSTM katmanlarının değiştirilmesi ile test edilmiştir.35] LSTM ile [36]



Figür 3. Bi-LSTM tabanlı yazım hatası tespit modelinin mimarisi.

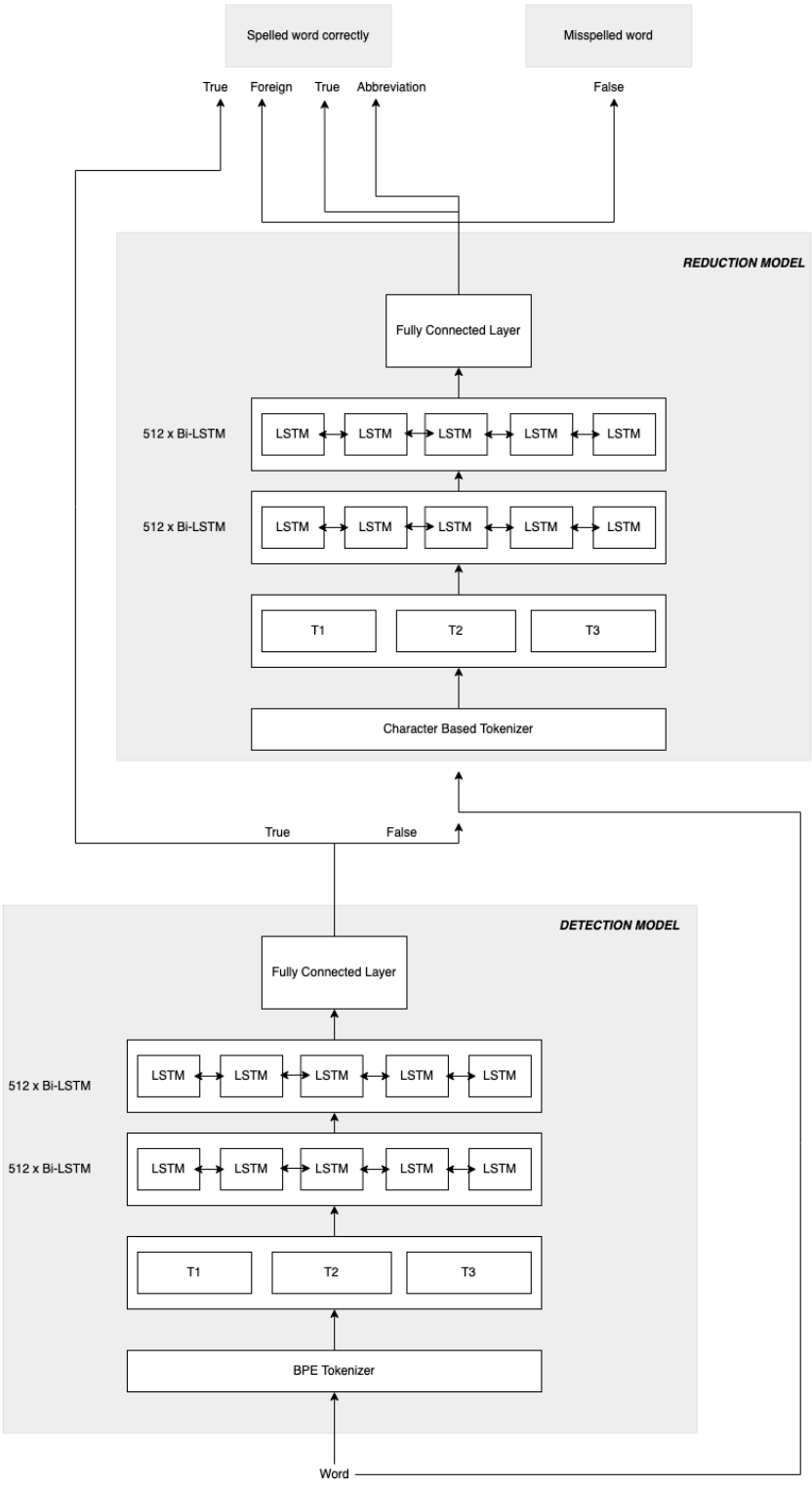
4.3. Yanlış pozitif azaltma modeli

Taban tespit modeli Türkçe kelimeler kullanılarak eğitilmiş ve Tabloda verilen fonksiyonlar kullanılarak bu kelimelerin çarpık versiyonları elde edilmiştir.1. Bu kelimeler temel tespit modelinin eğitimi için sözlükten alındığı için bu şekilde oluşturulan derlem Türkçede sıklıkla kullanılan yabancı kelime ve kısaltmaları içermemektedir. Bu nedenle yabancı kelimeler ve kısaltmalar modeller tarafından hatalı kelime olarak etiketlenmekte ve bu da modellerin hata oranının artmasına neden olmaktadır. Bu çalışmada, bu tür tespitlerden kaynaklanan yanlış pozitif tahminlerin sayısını azaltmak için bir yanlış pozitif azaltma modeli eğitildi. Yanlış pozitif azaltma modelinin mimarisi, temel tespit modeliyle birlikte Şekilde gösterilmektedir.4.

Kısaltmalar ve yabancı kelimeler sözlüğe dayalı olarak oluşturduğumuz kelime havuzunda bulunmadığından yanlış pozitif indirgeme modeli yalnızca karakter tabanlı tokenizer ile eğitilmektedir. Bu kelimeler Türkçenin hece yapısına da uygun değildir. Şekilde görüldüğü gibi4karakter tabanlı belirteçleyicinin ardından iki gizli katman gelir. Her gizli katmanda 512 Bi-LSTM düğümü bulunur. Son katman, elde edilen gösterimleri dört sınıftan birine eşleyen tamamen bağlı bir katmandır.

4.4. Önerilen birleşik model

Şekilde gösterilen önerilen kelime denetleyici modeli4iki farklı modelin birleşiminden oluşuyor. Öncelikle kelimeler ilgili belirteçle belirteçlere ayrılarak, doğru yazılan ve yanlış yazılan kelimelerin belirlenmesi için ikili sınıflandırma problemini ele alan temel tespit modeli uygulanır. deneyler



Şekil 4. Önerilen yazım hatası tespit modelinin mimarisi

Bölümde ayrıntılı olarak açıklanmıştır5baz tespit modelinin, birleştirilmiş modelde de kullanılan 512 Bi-LSTM düğümü ve iki gizli katmanla en iyi sonuçları elde ettiğini gösterdi. Doğru olarak sınıflandırılan kelimeler yazılanlar veri kümesinden çıkarılır. Temel tespit modeli tarafından yanlış yazılmış olarak etiketlenen geri kalan kelimeler, Bölüm'de ayrıntılı olarak açıklanan yanlış pozitif azaltma modeline beslenir.4.3. Bu modelin amacı, ilk modelde hatalı olarak işaretlenen sözcükleri yeniden kontrol ederek yanlış pozitiflik oranını azaltmaktır. Son olarak model, üçü kelimenin yazım hatası içermediğini gösteren dört tahminden birini üretir.

5. Sonuçlar ve tartışma

Bu bölümde öncelikle önerilen modelin çeşitli tokenizerler ve sinir ağı mimarileri ile detaylı bir değerlendirmesini sunuyoruz. Daha sonra Türkçedeki yazım hatası tespit problemlerini ele alan daha önce yapılmış ilgili çalışmalarla karşılaştırmalı sonuçlar veriyoruz.

5.1. Önerilen modelin değerlendirilmesi

Önerilen model farklı tokenizer ve sinir ağı mimarileri kombinasyonları ile değerlendirilmektedir. Tokenizasyon katmanında farklı derlem boyutlarına sahip karakter tabanlı, hece tabanlı ve BPE belirteçleri kullandık. Bu katmanı, her biri bir, iki ve üç gizli katmana sahip olan LSTM ve Bi-LSTM olmak üzere iki farklı sinir ağı mimarisi izledi. Her gizli katman 512 düğümden oluşur.

Masa3sonuçları doğruluk ve F1 puanı ölçümleri açısından gösterir. Eğitim için LSTM mimarisini kullanan ilk deneyler, kelime parçası tabanlı tokenizasyonun hem doğru hem de yanlış yazılmış kelimelerin tespitinde daha yüksek bir başarı oranı sağladığını gösterdi. Tabloda görüldüğü gibi3Karakter bazlı tokenizasyon ile LSTM kullanılarak elde edilen 0,65 olan F1 puanı, hece bazlı tokenizasyon yapıldığında 0,71'e yükselmiştir. Bu nedenle Bi-LSTM mimarisi ile yapılan deneyler yalnızca hece tabanlı ve BPE tokenizasyon yöntemleri kullanılarak gerçekleştirildi.

Tablo 3. Farklı tokenizer ve sinir ağı mimarileri kombinasyonlarından oluşan test setinde elde edilen sonuçlar.

Model adı	Tokenlaştırıcı	Tokenizer külliyatı	Gizlenmiş katman	Doğru kelimeler kesinlik	Yanlış yazılmış kelimelerin doğruluğu	Etraflı kesinlik	F1 Gol
512 LSTM	Karakter	Tüm kelimeler	2	%90	%70	%86	0,65
512 LSTM	Hece	Tüm kelimeler	2	%92	%75	%88	0.71
512 LSTM	BPE	Doğru kelimeler	1	%95	%84	%93	0,8
512 LSTM	BPE	Doğru kelimeler	2	%95	%88	%94	0,8
512 LSTM	BPE	Doğru kelimeler	3	%96	%88	%94	0,84
512 Bi-LSTM	BPE	Tüm kelimeler	2	%96	%90	%94	0,86
512 Bi-LSTM	Hece	Doğru kelimeler	2	%92	%90	%91	0,79
512 Bi-LSTM	BPE	Doğru kelimeler	1	%96	%90	%95	0,86
512 Bi-LSTM	BPE	Doğru kelimeler	2	%96**	%92**	%95**	0,87**
512 Bi-LSTM	BPE	Doğru kelimeler	3	%96	%91	%95	0.87

Tabloda görüldüğü gibi3Tespit modelinin eğitimi için LSTM mimarisi Bi-LSTM mimarisi ile değiştirildiğinde, hece tabanlı tokenizasyon ile önerilen modelin başarı oranı önemli ölçüde arttı. Bu nedenle Bi-LSTM mimarisini BPE tokenizasyonuyla daha da eğittik. Sonuçlar, BPE ile tokenleştirmenin, heceye dayalı tokenizasyona göre daha yüksek doğruluk ve F1 puanı sağladığını gösterdi. Ayrıca BPE tabanlı tokenizasyon için kullanılan derlemi de değiştirdik ve tokenizer olarak yalnızca doğru kelimelerin kullanıldığını gözlemledik.

Derlem, tüm kelimeler kullanılarak oluşturulan tokenizer derlemesine kıyasla biraz daha yüksek doğruluk ve F1 puanı verdi. Tabloda görüldüğü gibi3%95'lik en yüksek doğruluk ve 0,87'lik F1 puanı, BPE ve tokenizasyon için yalnızca doğru kelimeler ve eğitim için 2 gizli katmana sahip Bi-LSTM mimarisi kullanılarak elde edildi. Şekilde5Bi-LSTM modellerinin hece tabanlı tokenizer ve BPE tokenizer ile elde edilen eğitim ve doğrulama setleri üzerindeki eğitim ve doğrulama kaybı grafikleri gösterilmektedir. Görüldüğü üzere BPE tokenizer hem eğitimde hem de eğitimde daha düşük kayıp değerleri vermiştir.

doğrulama kümeleri.

Şekilde6, tokenleştirme işlemi için BPE tokenizer'ı yalnızca doğru kelimelerle kullanmanın avantajı alternatifleri ise iki örnek Türkçe kelime üzerinden gösterilmiştir. İlk örnekte “izleyiciler” olarak yazılması gereken yanlış yazılan “izliyiceler” kelimesi, karakter bazlı ve hece bazlı modellerle doğru yazılışlı kelime olarak etiketlenmiştir. Bu yanlış olumsuz tahminin nedeni yanlış yazılan izliyiceler şeklinin harf ve hece kullanımı açısından dilin uyumuna uygun olmasıdır. Öte yandan Şekilde görüldüğü gibi6BPE tokenizer bir tür sözlük tabanlı kontrol gerçekleştirdiğinden bu kelimenin yazım hatası BPE tabanlı model tarafından başarıyla tespit edildi. İkinci örnekte ise “herkes” şeklinde yazılması gereken yanlış yazılan “herkez” kelimesi, karakter bazlı ve hece bazlı modellerle doğru yazılan kelime olarak etiketlenmiştir. Görüldüğü gibi Türkçede yaygın olarak görülen bu hata, tokenizer temelli model sayesinde başarılı bir şekilde tespit edilmiştir.

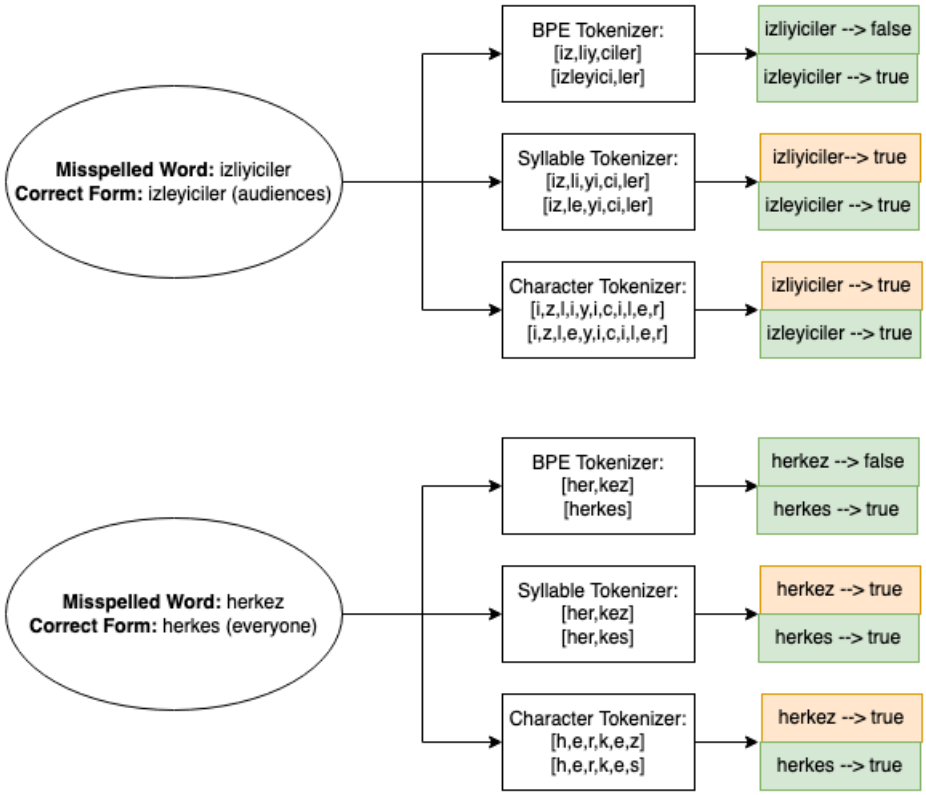
(a) Hece tabanlı belirteç (b) BPE tabanlı tokenizer

Şekil 5. Bi-LSTM modellerinin eğitim ve doğrulama setlerindeki kayıp grafikleri

5.2. İlgili çalışmalarla karşılaştırma

Bu bölümde önerilen modellerin Türkçe yazım hatası tespit yeteneğini ilgili çalışmalarla karşılaştırıyoruz. Ayrıca karşılaştırma amacıyla aynı veri kümesine iki son teknoloji transformatör tabanlı dil modeli olan BERT ve RoBERTa'yı (Sağlam Optimize Edilmiş BERT Ön Eğitim Yaklaşımı) uyguladık. Bu amaçla BERTürk'te[37] ve RoBERTaTurk[7] mimarileri sırasıyla BERT ve RoBERTa'ya dayalı önceden eğitilmiş Türkçe dil modelleri olup, Türkçe metindeki yazım hatalarını tespit etmek için bunların üzerine göreve özel katmanlar eklenmektedir. Bu modeller özel olarak Türkçe dil verileri üzerinde eğitilmiştir ve bu nedenle NLP görevlerini Türkçe metin üzerinde yüksek derecede doğrulukla gerçekleştirebilmektedirler.

Masa4kesinlik, geri çağırma, F1 puanı ve genel doğruluk ölçümleri açısından karşılaştırmalı sonuçları sunar. Önerilen modelin her iki versiyonunu da (üs tespit modeli ve birleştirilmiş model) uyguladık.



Şekil 6. Farklı belirteçler kullanan sinir ağı modelleri tarafından yanlış yazılan kelimelerin iki örneği için öngörülen etiketler

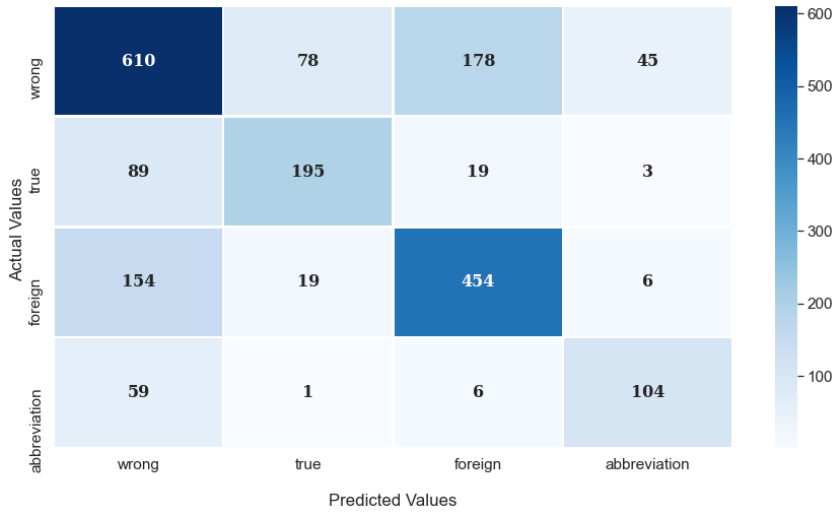
Bölümde detaylandırıldığı gibi4, birleştirilmiş model ek bir adım içerir; yanlış pozitif azaltma modeli, yalnızca temel yazım hatası tespit modeli tarafından yanlış yazılmış sözcükler olarak etiketlenen örneklere uygulanır. Modellerin Bölüm 1'de açıklanan aynı test setine uygulandığını da belirtmeliyiz.31974'ü doğru (olumsuz) ve 448'i yanlış yazılmış (pozitif) olmak üzere 2422 kelimeden oluşur.

Tabloda görüldüğü gibi4En yüksek F1 puanı ve genel doğruluk sırasıyla %0,91 ve %96,6 ile önerilen birleşik modelle elde edildi. Bu modelde de en yüksek hassasiyet elde edilirken, Microsoft Office yazılımındaki Türkçe yazım düzelticinin kullandığı modelin, nispeten çok düşük hassasiyetle daha yüksek hatırlama sağladığı gözlendi. Benzer şekilde Türkçede yaygın olarak kullanılan bir NLP kütüphanesi olan Zemberek de önerilen her iki modelden daha yüksek hatırlanma elde etmiştir. Ancak bu modelin hassasiyetinin tüm modellere göre daha düşük olduğu görülmektedir. Öte yandan sonuçlar, önerilen birleştirilmiş modelin en yüksek hassasiyete ulaşırken çok dengeli bir hassasiyet ve geri çağırma sağladığını gösterdi. İnce ayarlı transformatör tabanlı modeller BERTurk ve RoBERTaTurk sırasıyla 0,83 ve 0,84 F1 puanı verdi.

Tabloda görülen baz tespit modeli ile birleşik modelin karşılaştırılması4Yanlış pozitif azaltma modelinin uygulanmasıyla hatırlamada 0,3 puanlık bir azalmaya karşılık hassasiyette 0,9 puanlık bir artış elde edildiğini belirtti. Buna göre baz tespiti ile elde edilen F1 puanı ve genel doğruluk

Tablo 4. Test setindeki mevcut ve önerilen modellerin karşılaştırılması.

Yöntem	Kesinlik	Hatırlamak	F1 Puanı	Kesinlik
Zemberek20]	0,74	0,94	0.83	%92,8
Microsoft Office	0,78	**0,95**	0,86	%94,3
Eryiğit ve ark [23]	0,76	0.83	0,80	%92,1
Çolakoğlu ve ark.24]	0,89	0,86	0,88	%95,5
BERTürk [37]	0,84	0,74	0.83	%92,6
RoBERTaTürk [7]	0,85	0,75	0,84	%93,4
Hunspell [28]	0,84	0.71	0.82	%92,2
Önerilen Üs Tespit Modeli	0.82	0,94	0,88	%95,2
Önerilen Kombine Model	**0,91**	0,90	**0,91**	**%96,6**



Şekil 7. Test seti üzerindeki yanlış pozitif azaltma modelinin tahminleri.

model sırasıyla %0,3 ve %1,4 arttı. McNemar'ın testi, kombine modelin F1 puanı ile diğer tüm modeller arasındaki farkın istatistiksel olarak anlamlı olduğuna dikkat çekti. Yanlış pozitif azaltma modelinin daha ayrıntılı bir değerlendirmesini yapmak için, Bölüm'de ayrıntıları verilen mobil uygulamayı kullanarak etiketlenmiş numuneler üzerinde eğitim-test prosedürüyle ayırt edici yeteneğini test ettik.3. Veri setinin yüzde sekseni eğitim ve doğrulama amacıyla kullanılırken geri kalan örnekler test için kullanıldı. Şekilde verilen sonuçlar7Bu dört sınıflı problemde %67'lik genel doğruluğun elde edildiğini göstermektedir. Sonuçlar, modelin yabancı kelimeleri diğer sınıflardan ayırmadaki doğruluğunun %72 olduğunu ve bu oranın diğer sınıflar için elde edilen doğruluklardan daha yüksek olduğunu ortaya koymaktadır. Sınıf bazında en düşük doğruluk ise %61 ile kısaltma sınıfında elde edildi.

6. Sonuçlar

Bu çalışmada Türkçe için bir yazım hatası tespit modeli önerdik. Model iki ana adımdan oluşmaktadır. İlk adımda, verilen sözcüğe, doğru yazılan kelimelerle yanlış yazılan kelimeleri ayıran ikili sınıflandırma görevi için BPE belirteçli Bi-LSTM kullanılarak eğitilen bir model uygulanır. Saniyede

Adımda, pozitif yani yanlış yazılmış kelimeler olarak etiketlenen örnekler, yanlış pozitif tahminlerin azaltılmasının amaçlandığı yanlış pozitif azaltma modeline beslenir.

Yanlış pozitif azaltma modelinin eğitimi için geliştirilen bir mobil uygulama kullanılarak bir veri seti oluşturulmuş ve etiketlenmiştir. Bu çalışmanın bir katkısı olarak, etiketleme işlemi sırasında yabancı kelimeler ve kısaltmalar da tespit edilmiş ve böylece mevcut çalışmalardan farklı olarak önerilen birleşik model, sadece Türkçe kelimeleri değil aynı zamanda sosyal medya terimlerini içeren yabancı kelime ve kısaltmaları da tespit edebilmektedir. türkçe konuşma dilinde yaygın olarak kullanılmaktadır. Bulgular, yanlış pozitif azaltma modelinin, geri çağırmada nispeten daha az bir düşüş karşılığında baz tespit modelinin hassasiyetini önemli ölçüde arttırdığını gösterdi. Ayrıca, birleştirilmiş model, yazım hatası tespit problemlerini ele alan mevcut çalışmalara göre daha yüksek bir F1 puanı ve doğruluk sağlamıştır. Sonuçlar ayrıca, tokenleştirme için BPE tokenizer kullanmanın LSTM tabanlı modelin tespit yeteneğini geliştirdiğini de gösterdi.

Sosyal medya paylaşımları veya e-ticaret sitelerindeki ürün incelemeleri gibi çevrimiçi platformlardaki kullanıcı metinleri, yüksek oranda yanlış yazılmış kelimeler içermektedir. Model tabanlı metin düzeltme çalışmaları, model eğitimi için temiz bir metin veritabanı gerektirir. Gelecek araştırmalarda, hata tespit doğruluğu yüksek olan önerilen tespit modeli, derin öğrenme tabanlı bir yazım hatası düzeltme modeline temel oluşturacak temiz bir veri seti hazırlamak için kullanılabilir. Ayrıca, makalemizde kelime düzeyinde yazım algılama sorunlarını ele alırken, transformatör tabanlı mimarilerin, dikkat mekanizmaları ve üstün temsil etme yetenekleri nedeniyle cümle tabanlı bir yazım algılama görevini doğru bir şekilde modellemede LSTM modellerinden daha iyi performans göstermesinin beklendiğini belirtmekte fayda var. bağlamsal bilgi. Gelecekte, Türkçe için cümle tabanlı bir yazım tespiti ve düzeltme yaklaşımı geliştirerek araştırmamızı genişletmeyi ve bu makalede önerilen model ve dönüştürücü tabanlı modellerle kapsamlı bir karşılaştırma sağlamayı hedefliyoruz.

Referanslar

[1] Rapp A, Curti L, Boldi A. İnsan-sohbet robotu etkileşiminin insani tarafı: Metin tabanlı sohbet robotları üzerine on yıllık araştırmanın sistematik bir literatür taraması. Uluslararası İnsan-Bilgisayar Çalışmaları Dergisi. 2021;151:102630.3. https://doi.org/10.1016/j.ijhcs.2021.102630

[2] Singh D, Reddy S, Hamilton W, Dyer C, Yogatama D. Açık alan soru yanıtlama için çoklu belge okuyucu ve geri çağırıcının uçtan uca eğitimi. Sinirsel Bilgi İşleme Sistemlerindeki Gelişmeler. 2021; 34:25968-81.5

[3] Ali A, Nakov P, Bell P, Renals S. WERd: Diyalektik konuşma tanımayı değerlendirmek için sosyal metin yazım değişkenlerinin kullanılması. İçinde: 2017 IEEE Otomatik Konuşma Tanıma ve Anlama Çalıştayı (ASRU). IEEE; 2017; 7141-7148. https://doi.org/10.1109/ASRU.2017.8268928

[4] Bellegarda JR, Monz C. Dil ve konuşma işlemeye yönelik istatistiksel yöntemlerde son teknoloji. Bilgisayar Konuşması ve Dili. 2016; 35:163-84.10. https://doi.org/10.1016/j.csl.2015.07.001

[5] Birjali M, Kasri M, Beni-Hssane A. Duygu analizi üzerine kapsamlı bir araştırma: Yaklaşımlar, zorluklar ve eğilimler. Bilgiye Dayalı Sistemler. 2021; 226:107134.12. https://doi.org/10.1016/j.knosys.2021.107134

[6] Altuner AB, Kilimci ZH. Bilgi grafiğini ve topluluk bilincini kullanan yeni bir derin takviye öğrenmeye dayalı hisse senedi fiyatı tahmini. Türk Elektrik Mühendisliği ve Bilgisayar Bilimleri Dergisi. 2022;30(4):1506- 24.14. https://doi.org/10.55730/1300-0632.3862

[7] Aytan B, Sakar CO. Türkçe ve Farklı Dillerde Eğitilen Transformatör Tabanlı Modellerin Türkçe Doğal Dil İşleme Sorunlarında Karşılaştırılması. İçinde: 2022 30. Sinyal İşleme ve İletişim Uygulamaları Konferansı (SIU). IEEE; 2022. s. 1-4.17 (Türkçe ve İngilizce özet). https://doi.org/10.1109/SIU55565.2022.9864818

[8] Rivera-Trigueros I. Makine çeviri sistemleri ve kalite değerlendirmesi: sistematik bir inceleme. Dil Kaynakları ve Değerlendirme. 2021:1-27.19. https://doi.org/10.1007/s10579-021-09537-5

[9] Aggarwal CC. Metin için makine öğrenimi: Giriş. İçinde: Metin için makine öğrenimi. İlkbahar 2018; 1-16.https://doi.org/10.1007/978-3-319-73531-3_1

[10] Varma R, Verma Y, Vijayvargiya P, Churi PP. COVID-19 salgını öncesi ve sonrası sahte haber tespitine yönelik derin öğrenme ve makine öğrenimi yaklaşımları üzerine sistematik bir araştırma. Uluslararası Akıllı Bilgisayar ve Sibernetik Dergisi. 2021; 23. https://doi.org/10.1108/IJICC-04-2021-0069

[11] Yıldız B, Emekci F. Sosyal ağlar için ad yazım denetimi çerçevesi. Türk Elektrik Mühendisliği ve Bilgisayar Bilimleri Dergisi. 2016;24(4):2194-204.25. https://doi.org/10.3906/elk-1402-92

[12] Anbukkarasi S, Varadhaganapathy S. Doğal dil işlemede sinir ağı tabanlı hata işleyici. Sinirsel Hesaplama ve Uygulamaları. 2022;34(23):20629-38. https://doi.org/10.1007/s00521-022-07489-7

[13] Solak A, Oflazer K. Türkçe için yazım denetimi tasarımı ve uygulaması. Edebi ve dilsel hesaplama. 1993;8(3):113-30. https://doi.org/10.1093/llc/8.3.113

[14] Yessenbayev Z, Kozhirbayev Z, Makazhanov A. KazNLP: Kazak dilinde yazılmış metinlerin otomatik olarak işlenmesi için bir boru hattı. Uluslararası Konuşma ve Bilgisayar Konferansı 2020'de; 657-666. Springer, Cham. https://doi.org/10.1007/978-3-030-60276-5_63

[15] Özer H, Korkmaz EE. Transmorf: Türkçe için dönüştürücü tabanlı bir morfolojik belirsizliği giderici. Türk Elektrik Mühendisliği ve Bilgisayar Bilimleri Dergisi. 2022;30(5):1897-913. https://doi.org/10.55730/1300-0632.3912

[16] Choudhury M, Thomas M, Mukherjee A, Basu A, Ganguly N. Mükemmel bir yazım denetleyici geliştirmek ne kadar zor? Karmaşık ağ yaklaşımı yoluyla diller arası bir analiz. arXiv ön baskı fiziği/0703198. 21 Mart 2007. https://doi.org/10.48550/arXiv.physics/0703198

[17] Singh S, Singh S. HINDIA: Hintçe dilinin yazım denetimi için derin öğrenmeye dayalı bir model. Sinirsel Hesaplama ve Uygulamaları. 2021;33(8):3825-40. https://doi.org/10.1007/s00521-020-05207-9

[18] Hassan H, Menezes A. Bağlamsal grafik rastgele yürüyüşlerini kullanarak sosyal metin normalleştirme. Hesaplamalı Dilbilim Derneği'nin 51. Yıllık Toplantısı Bildirilerinde (Cilt 1: Uzun Yazılar) 2013;(s. 1577-1586).

[19] Oflazer K. Sondan eklemeli dillerde yazım düzeltme. arXiv ön baskı cmp-lg/9410004. 1994. https://doi.org/10.48550/arXiv.cmp-lg/9410004

[20] Akın AA, Dr. Akın. Zemberek, Türk dilleri için açık kaynaklı bir NLP çerçevesi. Yapı. 2007;10 (2007):1-5.

[21] Balik M. Yönlendirilmiş çevrimsel olmayan kelime grafiğinin uygulanması. Kybernetika. 2002;38(1):91-103.

[22] Damerau FJ. Yazım hatalarının bilgisayarla tespiti ve düzeltilmesi için bir teknik. ACM'nin iletişimi. 1964;7(3):171-6. https://doi.org/10.1145/363958.363994

[23] Eryiğit G, Torunoğlu-Selamet Di. Türkçe için sosyal medya metinlerinin normalleştirilmesi. Doğal Dil Mühendisliği. 2017;23(6):835-75. https://doi.org/10.1017/S1351324917000134

[24] Çolakoğlu T, Sulubacak U, Tantuğ AC. Kanonik olmayan Türkçe metinlerin makine çevirisi yaklaşımları kullanılarak normalleştirilmesi. Hesaplamalı Dilbilim Derneği'nin 57. Yıllık Toplantısında 2019 28 Temmuz. Hesaplamalı Dilbilim Derneği.

[25] Heafield K. KenLM: Daha hızlı ve daha küçük dil modeli sorguları. İstatistiksel makine çevirisi 2011 altıncı çalıştayının bildirilerinde (s. 187-197).

[26] Klein G, Kim Y, Deng Y, Senellart J, Rush AM. Opennmt: Sinirsel makine çevirisi için açık kaynaklı araç seti. arXiv ön baskı arXiv:1701.02810. 2017. https://doi.org/10.48550/arXiv.1701.02810

[27] Büyük O. Bağlama bağlı diziden diziye türkçe yazım düzeltmesi. Asya ve Düşük Kaynaklı Dilde Bilgi İşleme (TALLIP) ile ilgili ACM İşlemleri. 2020;19 (4):1-6. https://doi.org/10.1145/3383200

[28] Safaya A, Kurtuluş E, Göktoğan A, Yuret D. Mukayese: Turkish NLP Strikes Back. arXiv ön baskı arXiv:2203.01215. 2 Mart 2022. https://doi.org/10.48550/arXiv.2203.01215

[29] Viktor T, Gyorgy G, Halácsy P, Kornai A, Laszlo N ve ark. Hunmorph: açık kaynak kelime analizi.InWorkshop on Software 2005; 16:77-85

[30] Demir S, Topçu B. Grafik tabanlı Türkçe metin normalleştirmesi ve gürültülü metin işlemeye etkisi. Mühendislik Bilimi ve Teknolojisi, Uluslararası Bir Dergi. 2022; 35:101192. https://doi.org/10.1016/j.jetch.2022.101192

[31] Webster JJ, Kit C. NLP'de başlangıç aşaması olarak tokenizasyon. 14. Hesaplamalı Dilbilim Konferansı Bildirileri - Cilt 4 1992: s. 1106-1110. https://doi.org/10.3115/992424.992434

[32] Sennrich R, Haddow B, Birch A. Nadir kelimelerin alt kelime birimleriyle sinirsel makine çevirisi. arXiv ön baskı arXiv:1508.07909. 2015. https://doi.org/10.48550/arXiv.1508.07909

[33] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L ve ark. İhtiyacınız olan tek şey dikkat. Sinirsel bilgi işleme sistemlerindeki gelişmeler. 2017;30. https://doi.org/10.48550/arXiv.1706.03762

[34] Raffel C, Shazeer N, Roberts A, Lee K, Narang S ve ark. Birleşik bir metinden metne dönüştürücüyle aktarım öğreniminin sınırlarını keşfetme. Makine Öğrenimi Araştırma Dergisi. 2020;21 (140):1-67.

[35] Zhang S, Zheng D, Hu X, Yang M. İlişki sınıflandırması için çift yönlü uzun kısa süreli bellek ağları. 29. Pasifik Asya Dil, Bilgi ve Hesaplama Konferansı Bildirileri 2015 Ekim'de (s. 73-78).

[36] Hochreiter S, Schmidhuber J. Uzun kısa süreli hafıza. Sinirsel hesaplama. 1997;9(8):1735-80. https://doi.org/10.1162/neco.1997.9.8.1735

[37] Türkçe için Schweter S. Berturk-bert modelleri. 2020;3770924. https://doi.org/10.5281/zenodo.30