Deep learning-based Turkish spelling error detection with a multi-class false positive reduction model

Hazırlayan

Fırat Kaan BİTMEZ

Öğrenci Numarası

23281855

Dersin Hocası

Asst.Prof.Dr. İsmail İŞERİ

Özet

Bu yazıda OMÜ Doğal Dil İşleme Dersi kapsamında Ödev ve Öğrenme Amacıyla Tubitak'ta yayınlanan "Deep learning-based Turkish spelling error detection with a multiclass false positive reduction model" isimli makalenin Kısa bir özetini çıkarmak bu özeti yaparken de Makalenin Amacını Giriş, Gelişme ve Sonuç mantığıyla arayıp özetleyip Türkçe olarak sunmaktır. Orijinal Makale Yazarları:

Burak AYTAN, Cemal Okan SAKAR

Giriş

Asıl makalenin amacını bulmak için abstract ve giriş bölümden yaptığım çıkarımlar sonucunda bu makalenin konusu Türkçe'de dilin yazım hatalarının tespiti ve düzeltilmesinin metin normalleştirme sürecinde önemli bir adım olduğu Türkçe gibi sondan eklemeli bir dilde kök kelimenin birçok ek ile türetilmesi nedeniyle bu tür dillerde yazım hatalarını tespit etmenin zor olduğundan bahsedilmiş. Bu zorlukları aşmak için derin öğrenme ile bir model önerilmiş.

Makale Amacı ve Gelişme (Problem Tespiti)

Türkçe'de yazım hatalarını tespit etmek için derin öğrenmeye dayalı bir modelin geliştirilmesi ve yanlış pozitif tespitlerini azaltmak için çoklu sınıf yanlış pozitif azaltma modeli ile Türkçe metinlerde yazım hatalarını yüksek doğrulukla tespit etmek ve bu aşamaları gerçekleştirirken özel oluşabilecek durumları dikkate alarak etkili bir model oluşturmayı hedeflemektedir. Yazım hataları genel olarak sözcük dışı hatalar ve gerçek sözcük hataları olmak üzere iki kategoriye ayrılır Yanlış yazılan metin dilde var olan bir kelimeyse buna gerçek kelime hatası aksi halde kelime dışı hata adı verilir. Türkçe için yüksek doğruluklu bir yazım tespit modeli önermek için derin öğrenmeye dayalı bir model oluşturmak, Türk internet topluluğunda yaygın olarak kullanılan yabancı dil ve kısaltmalı içeren etiketli veri seti oluşturmak ve bunu mobil uygulama olarak sunmayı amaçlıyorlar. Çalışma aşamasında Solak ve Oflazer isimli kişilerin üç adımdan oluşan kural tabanlı bir yöntem önerisi inceleniyor. Bu yöntemin adımları kök belirleme, morfofonemik kontroller ve morfolojik ayrıştırmadır. Ana fikir ilk önce maksimum eşleşmeye dayalı bir algoritmaya dayalı olarak metnin kökünü bulmaktır. Bu algoritmada öncelikle kelimenin tamamı aranır.

Modeli oluşturmak için yapılan deneylerde kullanılan Veri Kümesinin Açıklanması: Tespit modelinin eğitimi ve temsil edilmesi için farklı kaynaklardan çeşitli veri setleri kullanılmış. Doğru yazılan kelimelerin yer aldığı veri seti TDK'dan alınan sözlük kullanılarak oluşturulmuş

1 milyon doğru ve 7 milyon yanlış kelime kullanılarak oluşturulan veri seti ana tespit modeli ve bu veri seti kullanılarak oluşturulmuştur. Model eğitimi için yanlış yazılan sözcüklerin oluşturulmasın yönelik işlevlere örnek olarak Sesli Harfleri kaldırmaya örnek olarak "merhaba">"merhba gibi yazılabilir bu yazım hatasından doğal dil işleme bu kelimeyi anlayamayabilir.

Yanlış yazılan kelimeler olarak sınıflandırılan örnekler çalışmasında olumlu tahminler olarak değerlendirilmiş. Yanlış Pozitif azaltma modeli yanlış yazılmış olarak etiketlenen doğru yazılmış sözcükler olan temel modellerin yanlış pozitif tahminlerini dahada azaltmak için tasarlanmış. Sistemin algılama yeteneğinin artırılması amacıyla kelime

bazlı etiketlemeye yönelik geliştirilen mobil uygulamada Labeled Word gönderiliyor Server'da New Word olarak geri döndürülüyor.

Tokenlaştırıcılar(Tokenizers) bölümünde bir cümleyi parağrafı yada metni kelimelere veya daha küçük parçaya ayırma işlemidir.

Baz Algılama modeli (Base detection model) bu modele girdi olarak bir kelimeyi alır ve kelimenin doğru yazılıp yazılmadığını çıktı olarak verir.

Yanlış Pozitif azaltma modeli kelimeler temel tespit modelinin eğitimi için sözlükten alındığı için bu şekilde oluşturulan derlemelerde Türkçede sıklıkla kullanılan yabancı kelimeler içermemetedir.Bu nedenle yabancı kelimeler kısaltmalar modeller tarafından hatalı kelime olarak etiketlenmekte ve bu da modelerin hata oranının artmasına neden olmaktadır.

Makalede Önerilen Birleşik Model iki farklı modelin birleşiminden oluşuyor. Öncelikle kelimeler ilgili belirteçle, belirteçlenerek ayırılıyor ve doğru yazılan , yanlış yazılan kelimelerin belirlenmesi için ikili sınıflandırma problemini ele alan tespit modeli uygulanıyor. Baz tespit modelinnin birleştirilmiş modelde kullanılan 512 Bi-LSTM düğümü ve iki gizli katmanla en iyi sonuçları elde ettiği gösterilmiş.

Sonuç

Bu makalede bahsedilen modeli oluşturmak için yapılan çalışmada yazım hatalarını tespit etmek ve düzetlmek için derin öğrenme tabanlı bir model üzerinden çalışılmış ve Bu uygulama sırasında kelime yanlış yazımlarındaki sorunlar ele alınmıştır. Bu sorunu çözerken kök kelimelerin birçok ek ile türetilmesi nedeniyle yazım hatalarının tespitinin zor olulduğu belirtilmiş Sosyal Medya sitelerindeki(Örneğin Twitter) elde edilen kullanıcı yazılarından el ile taglanmış (etiketlenmiş) kelimelerle oluşturulan test veri seti kullanılmış. Bu veriler üzerinde temel bir tespit modeli ve yanlış pozitif azaltma modeli geliştirilmiş. Ayrıca tokenizer kullanmının Türkçe yazım hatalarındaki tespitlerinde oluşan problemlerin etkisi incelenmiş ve transformer tabanlı dil modelleriyle karşılaştırılmıştır. Ve sonuçlar sunulmuştur.

Önerilen model, Türkçe metinlerde yazım hatalarını yüksek doğrulukla tespit etmeyi hedeflemektedir. Bu amaçla, TDK'dan alınan sözlük kullanılarak doğru yazılmış ve yanlış yazılmış kelimeler içeren bir veri seti oluşturulmuştur. Yanlış pozitif azaltma modeli, temel tespit modelinin yanlış pozitif tahminlerini azaltmak için tasarlanmıştır.

Önerilen birleşik model, belirli bir belirteçle ayırt edilmiş kelimeleri ikili sınıflandırma problemini ele alarak doğru ve yanlış yazılmış kelimeleri belirlemektedir. Bu modelin en iyi sonuçları, 512 Bi-LSTM düğümü ve iki gizli katman kullanılarak elde edilmiştir.

Sonuç olarak, Türkçe metinlerdeki yazım hatalarını tespit etmek ve düzeltmek için derin öğrenme tabanlı bir model önerilmektedir. Bu modelin, Türk internet topluluğunda yaygın olarak kullanılan yabancı dil ve kısaltmaları içeren etiketli veri setleri ile eğitilmesi ve mobil uygulama olarak sunulması amaçlanmaktadır. Bu çalışma, Türkçe metinlerdeki yazım hatalarının tespiti alanında önemli bir adım olabilir.

Kaynaklar

https://journals.tubitak.gov.tr/elektrik/vol31/iss3/7/